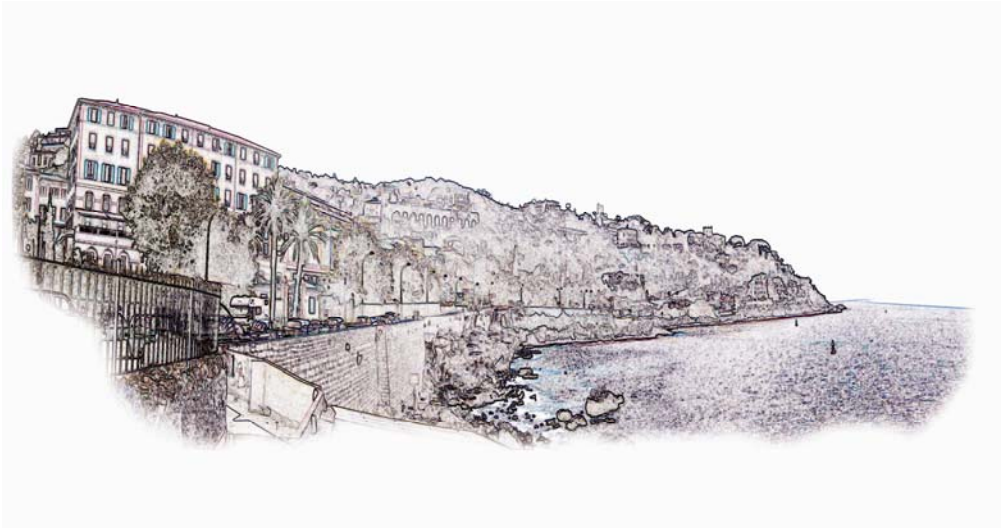
Plate-forme AFIA / Nice, du 30 mai au 3 juin 2005

**ATELIER :**

# Connaissance et Documents Temporels



Yannick Prié
Raphaël Troncy

# Organisateurs


– **Raphaël Troncy**
ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
email : raphael.troncy@isti.cnr.it
http://nmis.isti.cnr.it/troncy/

– **Yannick Prié**
LIRIS, UMR 5205 CNRS - Université Lyon 1,
69622 Villeurbanne Cedex, France
email : yannick.prie@liris.univ-lyon1.fr
http://liris.cnrs.fr/yannick.prie/


# Comité de Programme

– Bruno Bachimont (INA/UTC, Paris)
– Pierre-Antoine Champin (LIRIS, Lyon1)
– Patrick Gros (IRISA, Rennes)
– Julien Pinquier (IRIT, Toulouse)
– Yannick Prié (LIRIS, Lyon1)
– Cécile Roisin (INRIA Rhône-Alpes, Grenoble)
– Raphaël Troncy (ISTI/CNR, Pise)

# Avant-propos

Nous appelons *documents temporels* les documents multimédias dont au moins une des composante est temporelle, comme la vidéo, les flux sonores ou les documents hypermédias non statiques. De façon générale, l'instrumentation numérique des documents consiste à les intégrer dans un univers informatique qui en permette une exploitation d'abord équivalente à celle qui était possible hors de l'ordinateur, puis qui la dépasse. Cette instrumentation se poursuit depuis plusieurs décennies pour les documents statiques (textes ou images fixes) et a mené aux hypertextes, au web et au futur web sémantique. La difficulté est cependant plus grande en ce qui concerne les documents temporels, pour lesquels des techniques originales doivent être imaginées.

Nous nous intéressons dans le cadre de cet atelier à *l'utilisation de connaissances au sens large pour la manipulation de documents temporels* , et nous avons identifié au moins trois grands champs d'applications pour exploiter, analyser, décrire ou publier ces documents à l'aide de connaissances :

- L'extraction automatique de descripteurs dans les documents temporels est aujourd'hui guidée par des connaissances, que ce soit pour paramétrer les algorithmes travaillant directement sur le signal, ou pour analyser et interpréter leurs résultats dans l'objectif de combler une partie du fossé sémantique.
- La description et l'indexation de documents temporels utilisent les langages de représentation des connaissances tels que ceux développés autour du Web Sémantique. Ceux-ci permettent d'exprimer des descriptions de documents de manière à rendre leurs contenus plus accessibles et exploitables par les machines, notamment à travers le calcul d'inférences.
- La génération et la publication de nouveaux documents temporels à partir de fragments documentaires préalablement décrits peuvent elles aussi faire appel à des connaissances (modèles de publication, contraintes du support de restitution ou de l'environnement... ).

L'article de J. Pinquier et R. André-Obecht s'inscrit dans la première problématique pour proposer d'utiliser des modèles permettant de repérer des parties d'émissions télévisées et de structurer automatiquement des flux audiovisuels. Celui de Z. Ibrahim, I. Ferrane et P. Joly vise, dans le même champ, à la découverte de relations sémantiques structurelles entre fragments télévisuels. Dans le cadre de la représentation des connaissances, A. Isaac et R. Troncy proposent d'utiliser des ontologies formalisées pour décrire la structure et le contenu de documents télévisuels, et présentent une application menée à l'INA faisant usage des inférences ainsi rendues possibles. Finalement, S. Laborie, J. Euzenat et N. Layaida s'intéressent à l'utilisation de connaissances pour l'adaptation de documents multimédias, en utilisant notamment le langage SMIL.

Nous remercions François Brémond pour sa conférence invitée; les membres du comité de programme pour leurs relectures efficaces; ainsi que le comité d'organisation de la Plateforme AFIA 2005.

Le 31 Mai 2005, Yannick Prié & Raphaël Troncy

# Table des matières

# Interprétation Automatique de séquences vidéos

François Brémond

INRIA Sophia Antipolis, Projet ORION
2004 route des lucioles - BP 93 FR-06902 Sophia Antipolis
francois.bremond@sophia.inria.fr
http://www-sop.inria.fr/orion/personnel/Francois.Bremond/

**Résumé** : Ces travaux présentent une approche d'interprétation de séquences vidéos pour la génération automatique d'annotations. Cette approche permet de détecter et de suivre des objets mobiles (e.g. personnes) dans des vidéos et de reconnaître des scénarios d'intérêt. L'interprétation vidéo met en jeu des techniques de vision par ordinateur, de vision cognitive, de représentation des connaissances et de techniques d'apprentissage. Cette approche a montré des résultats encourageants dans de nombreux domaines applicatifs tels que la surveillance de supermarchés, d'autoroutes, de métros, de trains, du tarmac d'aéroports et d'agences bancaires. Des limitations demeurent : hypothèses d'utilisation restrictives, robustesse des algorithmes de vision (segmentation, classification, suivi) et acquisition fastidieuse des modèles de scénario. Cet exposé montrera les performances et limitations de l'interprétation vidéo pour la génération automatique d'annotations associées aux vidéos. Cet exposé présentera les nouvelles tendances dans le domaine (e.g. utilisation d'ontologies) pour structurer la connaissance nécessaire à l'obtention de solutions opérationnelles.

**Mots-clés** : Interprétation d'images, reconnaissance de forme, reconnaissance de scénario, séquence d'images

## 1 Interprétation Vidéo

L'interprétation vidéo a pour objectif de générer automatiquement en temps réel des alarmes et en différé des annotations associées aux vidéos. Plus généralement, l'interprétation automatique d'images est une problématique difficile qui est la base de nombreux travaux en vision et aussi en intelligence artificielle. La difficulté dépend de la nature des entités à reconnaître et du type d'interprétation recherchée. Il est plus simple de reconnaître des objets statiques et rigides en environnement manufacturé, que des comportements dynamiques de plusieurs objets non-rigides en environnement naturel. La difficulté dépend également du type d'interprétation recherchée. Le problème peut être soit, simplement, d'étiqueter une entité bien déterminée que l'on peut mettre directement

en correspondance avec des modèles, soit de détecter les entités, de les étiqueter et de vérifier leur cohérence (spatiale, temporelle, structurelle, etc). Les résultats de l'interprétation peuvent être la reconnaissance d'objets physiques, d'événements, de situations ou de scénarios. L'interprétation de séquences d'images a pour objectif, pour ce qui nous concerne, de donner un sens à une scène décrivant des activités humaines, à partir d'images fournies par une caméra couleur, monoculaire et fixe. Cette interprétation de scène repose, en général, sur la coopération d'un module de traitement d'images, d'un module de suivi des objets mobiles et d'un module de reconnaissance du comportement des objets mobiles qui s'appuient sur une base de contexte. Il s'agit, pour le module de traitement d'images, de détecter les régions mobiles sur la séquence d'images. Le module de suivi associe les régions détectées afin de former et de suivre les objets mobiles. La tâche du module de reconnaissance des comportements consiste, grâce à des techniques d'intelligence artificielle, à identifier les objets suivis et à reconnaître leur comportement comme constitutif d'un ou plusieurs scénarios prédéfinis. Un point important dans l'interprétation vidéo est ainsi la représentation et la reconnaissance de scénarios.

## 2 Représentation des Scénarios

Un formalisme de représentation permet de définir un scénario comme composé d'états ou d'événements. Un état est une propriété spatio-temporelle définie à un instant donné ou sur un intervalle de temps alors qu'un événement est composé d'un ou plusieurs changements d'états entre au moins deux états successifs ou sur un intervalle de temps. De plus, un scénario peut-être primitif (changement d'état simple) ou composé (combinaison d'états et/ou d'événements) (cf. tableau 1). Il est constitué de trois éléments : les objets physiques contiennent une liste de personnes et d'objets réels intervenant dans le scénario, les composants contiennent une liste d'états et d'événements présents dans le scénario, et les contraintes contiennent une liste de relations entre les objets physiques et les événements. Les objets physiques sont des objets mobiles ou contextuels. Les objets mobiles sont généralement une personne, un véhicule ou un groupe de personnes. Les objets contextuels sont des zones prédéfinies (e.g. zone d'entrée, local sensible) ou du mobilier (e.g. guichets, chaises).

| | |
|---|---|
| **événement_composé** Attaque_de_banque_avec_1personne | |
| **objets_physiques :** | ( (p : personne), (z1 : derrière_guichet), (z2 : salle_sensible), (g : porte_salle_sensible) ) |
| **composants :** | ( c1 : événement_primitif changement_de_zone(p, z1, z2) ) |
| **contraintes :** | (g est ouverte) |

TAB. 1 – Événement composé avec une personne utilisant la primitive changement_de_zone : la personne p se déplace de la zone z1 vers la zone z2 et la porte est ouverte.

# 3 Algorithme de Reconnaissance de Scénarios

L'algorithme temps-réel de reconnaissance comporte plusieurs étapes :

- Dans un premier temps, l'algorithme reconnaît itérativement les états primitifs en sélectionnant un ensemble d'objets physiques et en vérifiant les contraintes atemporelles correspondantes jusqu'à ce que toutes les combinaisons d'objets physiques aient été testées.
- Ensuite, l'algorithme reconnaît les événements primitifs. Un événement primitif a deux états primitifs comme composants. Lors de la reconnaissance des états primitifs, des instances (événements partiellement reconnus) sont construites pour chaque événement se terminant par l'état primitif reconnu. Ces instances contiennent la liste des objets physiques et le dernier composant relatifs à l'état primitif reconnu. L'algorithme recherche ensuite si le premier composant de l'événement primitif correspond à un des états primitifs reconnus dans le passé. Si les deux composants vérifient les contraintes définies dans le modèle, alors l'événement primitif est reconnu.
- Enfin, l'algorithme reconnaît les états et les événements composés. La reconnaissance de ces états et événements implique une recherche exhaustive de toutes les combinaisons possibles des objets physiques et des composants. Pour limiter une explosion combinatoire, tous les états et événements composés sont découpés en des états et des événements contenant au plus deux composants pendant une phase de pré-compilation. La reconnaissance des états et des événements composés est ainsi ramenée à celle des événements primitifs.

# 4 Problèmes en Génération Automatique d'Annotations de Vidéos

L'interprétation vidéo a montré des résultats encourageants dans de nombreux domaines applicatifs tels que la surveillance de supermarchés, d'autoroutes, de métros, de trains, du tarmac d'aéroports, d'agences bancaires et plus récemment pour la surveillance médicale de personnes âgées ou d'enfants hospitalisés. Des limitations demeurent : hypothèses d'utilisation restrictives (caméras fixes et connaissances a priori sur la scène observée), robustesse des algorithmes de vision (segmentation, classification, suivi) et acquisition fastidieuse des modèles de scénario. Pour pallier au problème de l'acquisition des connaissances (modèles de scénario), deux pistes sont actuellement envisagées. Premièrement, l'utilisation d'interfaces homme/machine permet à l'expert de définir ces scénarios en étant guidé par une ontologie du domaine d'application composée d'états et d'événements vidéos. Deuxièmement, des techniques d'apprentissage permettent d'apprendre automatiquement des modèles de scénarios récurrents se déroulant dans la scène.

# Structuration audiovisuelle par composantes sonores primaires

Julien PINQUIER et Régine ANDRÉ-OBRECHT

Équipe SAMOVA - IRIT - UMR 5505 CNRS INP UPS
118, route de Narbonne - 31062 Toulouse cedex 04, FRANCE
{pinquier, obrecht}@irit.fr

**Résumé** : Le développement croissant des données numériques et l'explosion des accès multimédia à l'information, sont confrontés au manque d'outils automatiques efficaces. Dans ce cadre, plusieurs approches relatives à l'indexation et la structuration de la bande sonore de documents audiovisuels sont proposées. Leurs buts sont de détecter les composantes primaires telles que la parole, la musique et les sons clés (jingles, sons caractéristiques, mots clés...). Pour la classification parole/musique, quatre paramètres originaux sont extraits : la modulation de l'entropie, la modulation de l'énergie à quatre hertz, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Des expériences sur un corpus radiophonique montrent la robustesse de ces paramètres : notre système possède un taux de classification correcte supérieur à 90 %. Un autre partitionnement consiste à détecter des sons clés. La sélection de candidats potentiels est effectuée en comparant la « signature » de chacun des jingles au flux de données. Ce système est simple par sa mise en œuvre mais rapide et très efficace : sur un corpus audiovisuel d'une dizaine d'heures (environ 200 jingles) aucune fausse alarme n'est présente. Il y a seulement deux omissions dans des conditions extrêmes. Les applaudissements sont modélisés à l'aide de MMG dans le domaine spectral. Un corpus télévisuel permet de valider cette première étude par des résultats encourageants. Grâce à l'extraction de ces composantes primaires, les émissions audiovisuelles peuvent être annotées de manière automatique. Une réflexion est conduite quant à l'utilisation de ces composantes afin de trouver une structure temporelle à nos documents : il s'agit de détecter un motif récurrent dans une collection d'émissions, dites de plateau.

**Mots-clés** : indexation sonore, structuration audiovisuelle, classification, énergie, entropie, segmentation, parole, musique, jingles, applaudissements.

# 1 Introduction

Par analogie avec les documents textuels qui sont faciles à manipuler (stockage, manipulation et recherche d'information étant devenus des opérations abordables par le grand public), le traitement des documents multimédia n'en est qu'à son balbutiement. Par exemple, trouver la vidéo contenant les premiers pas d'Armstrong sur la Lune (sans information *a priori*) est pour l'instant assez critique si l'on ne traite que les documents multimédia. Il serait souhaitable, comme en indexation textuelle, que l'on puisse utiliser des moteurs de recherche via des mots clés. Cela nécessite d'extraire du sens de la vidéo et/ou de l'audio et de les utiliser conjointement.

Un document sonore, c'est-à-dire la bande sonore d'un document multimédia ou enregistrement d'émission radiophonique, est un document particulièrement difficile à indexer, car l'extraction de l'information élémentaire se heurte à l'extrême diversité des sources acoustiques. Les segments acoustiques sont de nature très diverses de par leur production et leur enregistrement : l'environnement peut être propre ou plus ou moins bruité, la qualité de l'enregistrement peut être plus ou moins soignée et liée à des éléments extérieurs (canal téléphonique), la musique peut être traditionnelle ou synthétique, la présence de parole peut être observée en monologue ou en dialogue...

Il peut être intéressant de rechercher des « bruits » ou des sons sémantiquement significatifs tels que les applaudissements, les rires ou les effets spéciaux (pistolets, explosions...), de repérer les passages musicaux pour les segmenter et les identifier, de détecter les locuteurs équivalents à des tours de parole dans un dialogue. Si l'on se réfère à la norme MPEG7, indexer un document sonore signifie rechercher aussi bien des composantes de bas niveau dites primaires comme la parole, la musique, les sons clés (jingles, mots-clés...) que des descripteurs de plus haut niveau tels les locuteurs ou les thèmes.

Dans cet article, nous présentons un système de détection de composantes primaires telle la parole et la musique. Ce système est fondé sur l'extraction de trois paramètres originaux : la modulation de l'entropie, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Les informations issues de ces paramètres sont ensuite fusionnées avec celle issue de la modulation de l'énergie à quatre hertz. Au delà du partitionnement primaire, il est intéressant de détecter des sons clés ou des jingles représentant le début et/ou la fin d'un segment sonore afin de structurer le flux audio-visuel (Carrive *et al.*, 2000). Il ne s'agit pas de rechercher des thèmes (Amaral *et al.*, 2001), mais plutôt de proposer une macrosegmentation de l'audio en trouvant sa structure temporelle. Nous décrivons nos systèmes de détection de jingles et d'applaudissements. Ensuite, pour chacun des systèmes, nous proposons des expériences sur des corpora radiophoniques et télévisuels. Enfin, nous présentons deux exemples de fusion possible de nos outils en vue d'une structuration de plus haut niveau : il s'agit d'une recherche de motif dans une collection d'émissions et d'une segmentation d'un journal télévisé.

# 2 Système de détection parole/musique

Le système se décompose en deux systèmes de classification correspondant aux deux détections disjointes de la parole et de la musique. Il est fondé sur l'extraction de quatre paramètres (cf. figure 1) : la modulation de l'énergie à 4 Hertz, la modulation de l'entropie, le nombre de segments par seconde et la durée de ces segments (Pinquier *et al.*, 2003). La décision est prise en comparant les scores (vraisemblances) issus de la modélisation de chacun des paramètres considérés.

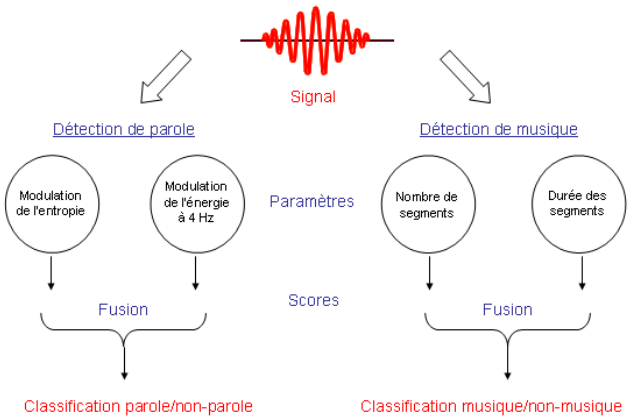



FIG. 1 – Le système de classification parole/musique.

## 2.1 Détection de parole

La détection des zones de parole est effectuée à partir de la fusion des deux paramètres de modulation.

### 2.1.1 Modulation de l'énergie à 4 Hertz

Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hertz (Houtgast & Steeneken, 1985). En effet, ces modulations correspondent au rythme syllabique. La parole possède une modulation de l'énergie à 4 Hertz plus forte que la musique.

### 2.1.2 Modulation de l'entropie

Des observations menées sur le signal ainsi que sur le spectrogramme font apparaître une structure plus « ordonnée » du signal de musique que de parole. Pour mesurer ce « désordre », nous avons calculé un paramètre fondé sur l'entropie du signal (Moddemeijer, 1989). La modulation de l'entropie est alors plus élevée pour la parole que pour la musique (cf. figure 2).

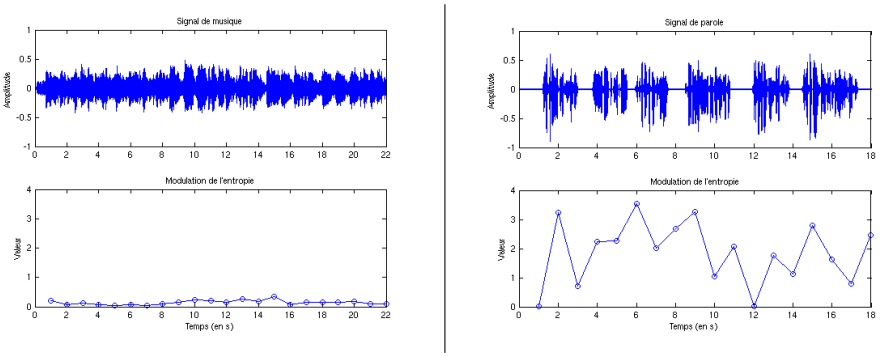



FIG. 2 – Modulation de l'entropie pour la musique (extrait de Mozart de 22 s) et la parole (6 phrases de parole lue de 18 s).

## 2.2 Détection de musique

La détection des zones de musique est réalisée grâce à deux paramètres issus d'une segmentation automatique du signal. La longueur des segments quasi stationnaires est différente pour la parole et la musique. En utilisant une segmentation du signal en zones quasi stationnaires, nous cherchons à mettre en évidence cette information. La segmentation est issue de l'algorithme de « Divergence Forward-Backward » (DFB) (André-Obrecht, 1988) qui est fondé sur une étude statistique du signal dans le domaine temporel.

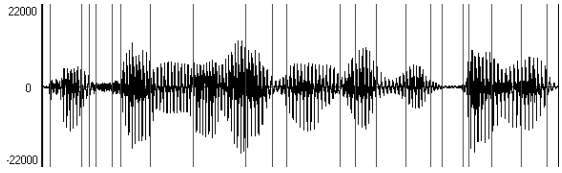



FIG. 3 – Résultat de la segmentation sur environ 1 seconde de parole. La phrase prononcée est : « Confirmez le rendez-vous par écrit ».

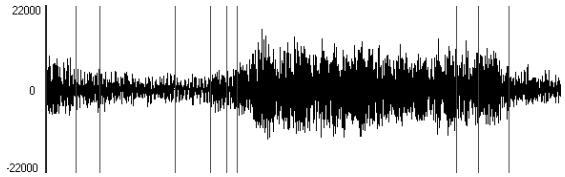



FIG. 4 – Résultat de la segmentation sur environ 1 seconde de musique d'un extrait de Mozart.

### 2.2.1 Nombre de segments

Le nombre de segments présents durant chaque seconde de signal est calculé. Les signaux de parole présentent une alternance de périodes de transition (voisées/non-voisées) et de périodes de relative stabilité (les voyelles en général) (Calliope, 1989). Au niveau de la segmentation, cela se traduit par de nombreux changements. La musique, étant plus tonale (ou harmonique), ne présente pas de telles variations. Le nombre de segments par unité de temps (ici la seconde) est donc plus important pour la parole que pour la musique.

### 2.2.2 Durée des segments

La durée des segments, obtenue après segmentation automatique (DFB), est fortement corrélée au nombre de segments par seconde. Afin de limiter la corrélation de ces deux paramètres de segmentation, la durée moyenne des segments sur une seconde est calculée sur les 7 segments les plus longs de la seconde. Le nombre de segments caractéristiques est fixé expérimentalement. Les segments sont généralement plus longs pour la musique que pour la parole.

## 3 Système de détection de jingles

Un jingle est un extrait sonore qui dure généralement quelques secondes. Il a pour but de présenter le début ou la fin d'une émission (météo, journal, publicité...) ou d'attirer l'attention de l'auditeur. Il a la particularité de pouvoir aussi bien contenir de la musique que de la parole. Il est, de plus, généralement redondant dans une collection de documents audiovisuels. Le système de détection d'un jingle est divisé en trois modules classiquement utilisés dans un problème de reconnaissance de formes (cf. figure 5).

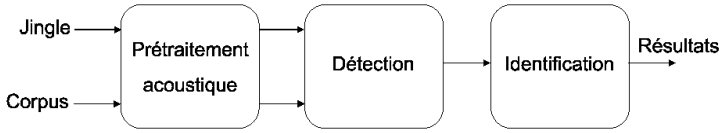



FIG. 5 – Schéma général de détection d'un jingle.

## 3.1 Prétraitement acoustique

Le pré-traitement acoustique est fondé sur une analyse spectrale. Le signal est découpé en trames de 32 ms avec recouvrement sur la moitié de la trame. 28 coefficients spectraux sont extraits (Pinquier *et al.*, 2002).

## 3.2 Détection et identification

Un jingle de référence est caractérisé par une suite de $N$ vecteurs spectraux que nous appelons « signature ». Cette valeur correspond au nombre de fenêtres d'analyse obtenues sur la durée totale du jingle considéré. La détection consiste à trouver cette séquence (suite de vecteurs) dans le flux de données à analyser. La distance Euclidienne est utilisée afin de comparer la signature du jingle et le signal. Cette comparaison s'effectue donc sur une fenêtre glissante de vecteurs que l'on déplace par un pas de $S$ vecteurs (cf. figure 6).

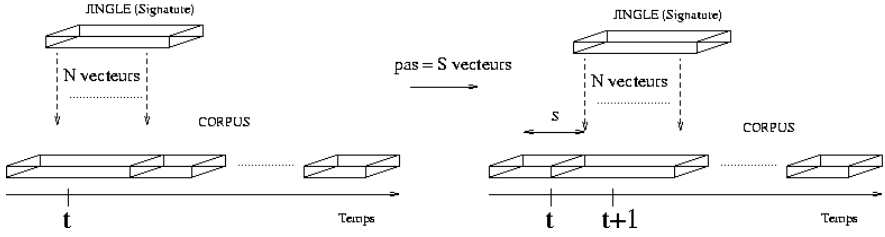



FIG. 6 – Comparaison entre le jingle et le corpus par distance Euclidienne.

Les candidats potentiels correspondent à certains minima locaux de la distance signature/flux calculée. L'analyse proposée consiste à étudier la largeur des pics de chacun des minima locaux à l'aide des variables de valeur du minimum local $h$, de hauteur relative $H$ et de largeur de pic $L$ (cf. figure 7).

Ce système est plus détaillé dans (Pinquier & André-Obrecht, 2004).

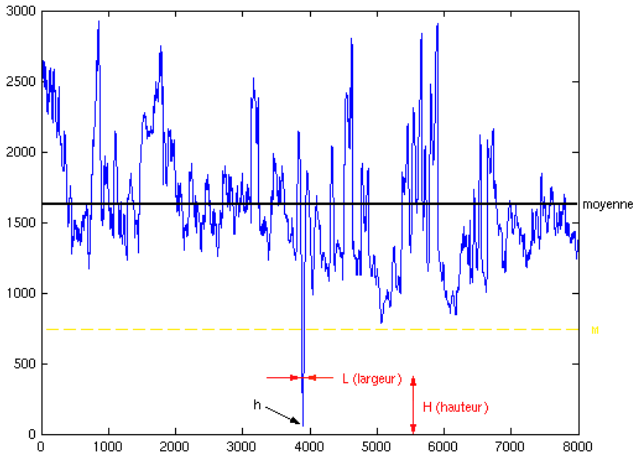



FIG. 7 – Sélection des jingles.

# 4 Détection des applaudissements

Le système est composé de deux modules principaux : le prétraitement et la décision (ou reconnaissance). Le prétraitement correspond à une analyse spectrale (voir section 3.1). Le système de détection mis en place s'inspire très largement des systèmes classiques parole/musique (Gauvain *et al.*, 1999). Il consiste à identifier sur chaque trame de signal, la présence ou l'absence du phénomène considéré en question (parole, musique, applaudissements...). Il s'agit d'un problème de classification en classe (applaudissements) et non-classe (non-applaudissements). La modélisation est effectuée à l'aide de Modèles de Mélanges de lois Gaussiennes et un apprentissage est alors nécessaire.

## 4.1 Reconnaissance

La décision se fait suivant la règle du maximum de vraisemblance entre les modèles applaudissements et non-applaudissements. Une fonction de lissage permet de ne garder que les segments significatifs (représentatifs d'une zone d'applaudissements), après regroupement des trames correspondant à la même décision. Le lissage est d'une seconde.

## 4.2 Apprentissage

L'apprentissage de ces paramètres est classiquement réalisé par les algorithmes VQ (Lloyd, 1957) pour l'initialisation des modèles et EM (Dempster *et al.*, 1977) pour la réestimation et l'optimisation des paramètres du mélange.

Après expérimentations, le nombre de lois gaussiennes de chacun des modèles a été fixé à 64 (cf. figure 8).

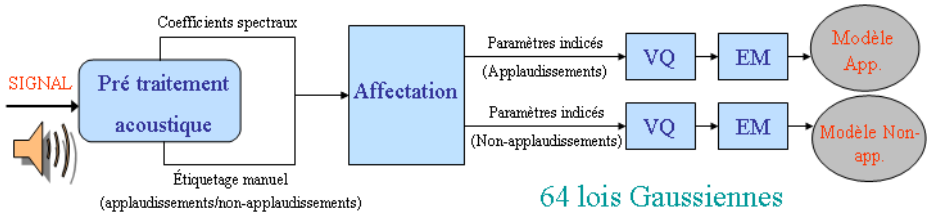



FIG. 8 – Apprentissage des modèles de mélanges de lois gaussiennes représentant les applaudissements et les non-applaudissements.

# 5 Expériences

## 5.1 Corpus

Notre base de données est assez hétérogène. L'apprentissage des seuils du système de classification parole/musique a été effectué sur 3 heures de parole lue : corpus MULTEXT (Campione & Véronis, 1998) et sur 3 heures d'extraits musicaux. Les tests sont effectués sur des données radiophoniques et télévisuelles pour une durée totale d'environ 12 heures.

Beaucoup de jingles apparaissent dans notre base de données. Notre but est de détecter et d'identifier seulement les jingles similaires à ceux de notre catalogue de sons clés. Nous avons 132 jingles à retrouver et reconnaître, sachant que notre catalogue est composé de 32 jingles de référence.

Pour la détection des applaudissements, nous utilisons le corpus « Le Grand Échiquier ». Le contenu de ce corpus est assez divers : de la musique (classique, jazz, variété française...), des interviews et des sketches. Chaque émission a une durée d'environ 190 minutes. La première que nous appelons « GE1 » nous sert d'apprentissage et la seconde « GE2 » nous sert de test. Nous utilisons donc au total seulement 6 heures de notre corpus. L'utilisation d'autres émissions est possible, des tests ont d'ailleurs été effectués sur deux autres émissions. La tâche d'annotation manuelle nécessaire pour évaluer les résultats étant très pénible, nous avons fait cet étiquetage que pour deux émissions seulement.

## 5.2 Evaluation

### 5.2.1 Classification parole/musique

Chaque paramètre est pertinent dans le sens où il permet, en lui-même, de faire une discrimination parole/non-parole ou musique/non-musique correcte. En considérant chaque paramètre individuellement, le taux de classification correcte varie d'environ 78 % pour la durée des segments à plus de 87 % pour la modulation de l'entropie (Table 1).

TAB. 1 – *Résultats de la classification parole/musique.*

| Paramètres | Taux de classification correcte |
|---|---|
| (1) Modulation de l'énergie à 4 Hz | 87.3 % |
| (2) Modulation de l'entropie | 87.5 % |
| (3) Nombre de segments | 86.4 % |
| (4) Durée des segments | 78.1 % |
| (1) + (2) Détection de parole | **90.5 %** |
| (3) + (4) Détection de musique | **89 %** |

La fusion entre ces paramètres par maximisation des scores de vraisemblances permet d'améliorer les résultats et d'obtenir environ 90 % de reconnaissance correcte pour chacune des classifications (parole/non-parole et musique/non-musique).

### 5.2.2 Détection de jingles

Sur les 132 jingles que nous devions localiser et identifier, nous en avons détecté 130, soit 98,5 % de taux de reconnaissance. Les deux seuls jingles omis (un jingle « France Info » et un jingle publicitaire) sont complètement recouverts de parole (le présentateur parle durant le jingle !) et leur pic est dans ce cas beaucoup trop large. La détection est excellente car nous n'avons aucune fausse alarme (insertion) et seulement deux omissions alors que d'autres jingles n'appartenant pas au catalogue de sons clés sont présents dans la base de données.

Durant la phase d'évaluation, nous avons étudié la précision de la détection. La localisation des jingles est très bonne : la différence entre les localisations manuelle et automatique est très faible, inférieure à 500 ms quel que soit le jingle ; elle correspond au pas $S$ utilisé.

### 5.2.3 Détection des applaudissements

Les résultats de la détection des applaudissements sont intéressants : nous obtenons 98,58 % de taux de classification correcte. Ce taux est à relativiser : il nous faut observer les applaudissements retrouvés. Sur les 906 secondes d'applaudissements que nous avons repérées manuellement, nous avons relevé 144 segments. Seulement 72 de ces segments sont significatifs et ceux-ci sont tous

bien détectés par notre système. Lorsque nous employons le terme « significatif », il désigne des *segments assez longs*, de durée supérieure à 1 seconde, et *pur* : ce ne sont pas des segments de faibles amplitudes ou superposés à de la parole. Le système de détection des applaudissements est excellent car il n'y a pratiquement pas d'insertions et tous les segments significatifs sont retrouvés.

# 6 Structuration

Nous nous plaçons ici dans le cadre d'une première analyse de scène par les composantes sonores primaires : la recherche de parole, de musique, de jingles et d'applaudissements. Ces informations de « bas niveau », extraites directement du flux sonore, ne sont pas directement exploitables pour la structuration de documents audiovisuels. Pour accéder à une information de plus haut niveau, il faut d'une part les regrouper, et d'autre part voir leurs impacts sur les autres informations sonores. Nous proposons un exemple de fusion possible de nos outils de segmentation sonore en vue d'une structuration de plus haut niveau : une recherche de motif sur une collection d'émissions.

## 6.1 Détection de motif dans une collection d'émissions

### 6.1.1 Présentation

Lorsqu'une seule émission est analysée, un traitement très spécifique peut être effectué. Suivant la durée de l'émission, un traitement manuel peut même être réalisé et s'avérer plus rapide. Par contre, quand nous sommes en présence de plusieurs émissions, comme la collection du « Grand Échiquier » présentée dans la section 5.1 dont nous possédons 54 émissions, il est nécessaire de changer de stratégie. Les traitements ne peuvent être qu'automatiques vue la durée du corpus (plus de 160 heures). Le niveau structurel doit rester assez grossier afin de correspondre à toute la collection d'émissions considérée.

L'étude de cette collection a permis de définir un motif, c'est-à-dire un enchaînement récurrent de caractéristiques communes à chacune des émissions. Ce motif structure les émissions en parties homogènes ; en général, il s'agit du passage d'un invité à un autre. Le motif est le suivant :
**présentateur** / *[applaudissements]* / **spectacle** / *[applaudissements / spectacle]* / **applaudissements** / **présentateur**.

Ceci signifie que dans cette collection, un spectacle (chanson, morceau de musique, sketch, extrait de film...) est introduit par le présentateur et est suivi par des applaudissements. À la fin de ceux-ci, le présentateur reprend la parole. Des applaudissements peuvent éventuellement précéder la composition artistique ou la « découper ».

Nous avons choisi de rechercher cette structuration sur le même fichier de test que précédemment pour des raisons évidentes : nous possédons déjà la « vérité terrain » correspondant aux détections d'applaudissements et du présentateur pour cette émission. Lors de cette émission, nous avons répertorié une succession de dix de ces motifs. Afin de retrouver le motif en question, de manière automatique, nous allons appliquer une stratégie dite « aveugle ».

Trois classifications sont effectuées indépendamment les unes des autres (cf. figure 9) :
- une détection de musique permettant de repérer les chansons,
- une détection de parole, pré-traitement pour la recherche du présentateur,
- une détection des applaudissements.

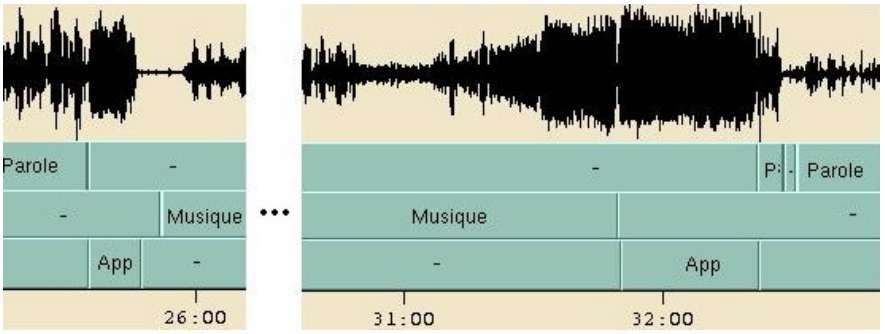



FIG. 9 – Exemple de recherche de motif sur 7 minutes de l'émission « GE2 » de la collection du « Grand Échiquier » à travers les détections automatiques de parole (ligne 1), de musique (ligne 2) et d'applaudissements (ligne 3).

Notons, que les modèles des applaudissements et des non-applaudissements sont les mêmes que ceux développés précédemment. Ils ont été appris sur une autre émission : « GE1 ». Rappelons aussi, que pour les détections de parole et de musique, notre système ne nécessite pas d'apprentissage. Il n'y a donc eu aucun apprentissage sur GE2.

La figure 10 est un exemple de résultat obtenu par une mise en commun de tous les résultats de détection.

Les résultats sont excellents sur l'émission « GE2 ». Les 10 motifs cherchés sont retrouvés et ceci bien que les détections de parole et de musique soient imparfaites : quelques légers décalages et des insertions de parole sur la musique sont observés.

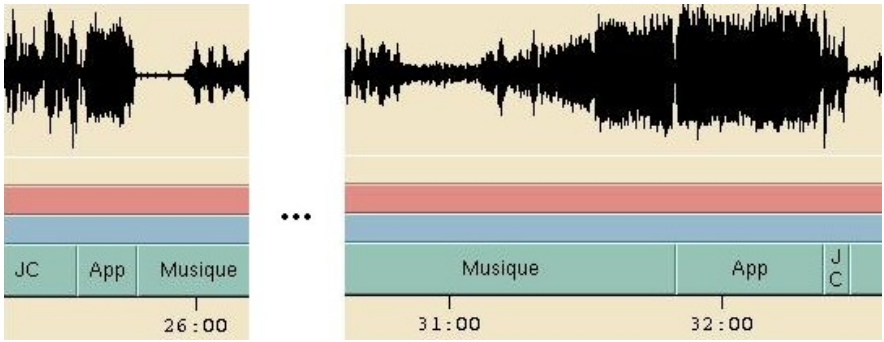



FIG. 10 – Exemple de résultat obtenu par fusion des différentes détections sur un extrait de 7 minutes de l'émission « GE2 » du « Grand Échiquier ». « JC » représente Jacques Chancel.

### 6.1.2 Remarques

Afin de valider totalement notre approche, l'ensemble de cette structuration va être effectuée sur l'ensemble des 54 émissions de cette collection sans aucun autre apprentissage ou intervention manuelle. Comme nous l'avons déjà signalé, le motif recherché n'est pas spécifique au « Grand Échiquier » mais commun à nombre d'émissions de plateaux.

La recherche d'un tel motif sur une autre émission ne nécessite alors qu'un étiquetage de 2 ou 3 minutes du nouveau présentateur de l'émission afin de créer son propre modèle acoustique. Ceci est effectué par adaptation du modèle dit du « monde » Les autres détections ne nécessitent pas d'apprentissage.

Le choix de ce motif s'est effectué après une intervention humaine, à savoir l'écoute de deux émissions du « Grand Échiquier ». Il s'est avéré que ce motif a pu être retrouvé tout au long des émissions et a été ainsi validé.

Il serait intéressant de pouvoir trouver de manière automatique le motif récurrent d'une émission. Le principe pourrait être le suivant :

- annotation automatique à partir des outils d'analyse audio et vidéo,
- recherche automatique des suites récurrentes dans la succession des annotations,
- inférence d'un motif,
- structuration du document à partir du motif trouvé.

La deuxième étape s'apparente à une détection d'invariants audiovisuels et fait l'objet de recherche récentes Haidar *et al.* (2004).

# 7 Conclusion

Dans le contexte de l'indexation sonore, nous avons étudié différentes composantes primaires, permettant une structuration audiovisuelle. Pour chacune de ces unités bas niveau, un détecteur automatique est développé afin de les extraire du continuum sonore.

Pour les spécialistes de l'audio, les composantes primaires correspondent souvent à la parole et à la musique. Le système original que nous avons développé est fondé sur une fusion de quatre paramètres : la modulation de l'énergie à 4 Hertz, la modulation de l'entropie, le nombre de segments issus d'une segmentation automatique et la durée de ces mêmes segments. Les résultats obtenus sont très bons, plus de 90 % de classification correcte, mais surtout le système est très robuste : il ne nécessite aucun nouvel apprentissage et/ou adaptation de ses seuils contrairement aux approches classiques fondées sur une analyse spectrale et des modèles de mélanges de lois gaussiennes.

D'autres composantes primaires correspondent à des sons clés : nous avons étudié les jingles et les applaudissements. Notre détecteur de jingles est excellent. Bien que la méthode soit assez simple, mesure de distance dans le domaine spectral, sur 10 heures de tests et 132 jingles à retrouver, nous n'avons observé que 2 omissions et aucune fausse alarme. Les erreurs apparaissent dans des conditions très particulières, par exemple là où la parole recouvre entièrement le jingle. Cette étude est d'autant plus intéressante qu'elle ne nécessite aucun apprentissage, seulement une occurrence du jingle à reconnaître. Une étude sur la détection des applaudissements a permis de montrer la faisabilité d'une méthode fondée sur une analyse spectrale et des modèles de mélanges de lois gaussiennes. Les résultats sont excellents car les sons caractéristiques, utiles en vue d'une structuration, sont tous détectés.

À la suite des détections de ces différentes composantes primaires, une étude en structuration a été effectuée : il s'agit d'une détection de motif sur une collection d'émissions. Elle permet de mettre en parallèle nos différentes détections afin d'extraire un enchaînement récurrent dans des émissions issues de la collection le « Grand Échiquier ». Il s'agit de la séquence : présentateur / [applaudissements] / spectacle / [applaudissements / spectacle] / applaudissements / présentateur. Durant notre émission de test, les 10 occurrences de ce motif sont toutes détectées.

Ces premiers travaux de structuration sont encourageants, mais il est fort dommage de se limiter à l'analyse d'un seul media (le son dans notre cas), alors que nous exploitons des bases de données audiovisuelles. C'est pourquoi nous devrons réfléchir sur l'apport de l'analyse vidéo. La détection de logos et la reconnaissance de l'intervenant sont des études complémentaires à la détection de jingles et à la reconnaissance de locuteur.

Ces travaux sont actuellement en cours au sein de notre équipe sur l'analyse de la vidéo.

Cette immersion dans l'analyse de la vidéo doit nous permettre de mieux cerner les complémentarités entre l'audio et la vidéo et d'appréhender à moyen terme le vrai problème du traitement audiovisuel. Il est impératif de savoir répondre plus précisément à des questions classiques du type :

- qu'est ce qu'une information audiovisuelle ? Qu'est ce qu'une indexation audiovisuelle ? La présence d'un personnage est une illustration simple de ce type d'information, voire d'index. Peut-on généraliser cette démarche ?
- qu'est ce qu'une analyse audiovisuelle ? Sachant que, comme nous avons essayé de le montrer, une analyse audiovisuelle ne signifie pas une simple fusion d'informations issues d'une analyse audio et d'une analyse vidéo.

## Références


Amaral R., Langlois T., Meinedo H., Neto J., Souto N. & Trancoso I. (2001). The Development of a Portuguese Version of a Media Watch System. In *European Conference on Speech Communication and Technology*, volume 4, p. 2689–2692, Aalborg, Denmark.

André-Obrecht R. (1988). A New Statistical Approach for Automatic Speech Segmentation. *IEEE Transactions on Audio, Speech, and Signal Processing*, **36**(1), 29–40.

Calliope (1989). *La parole et son traitement automatique*. Paris, France : Masson.

Campione E. & Véronis J. (1998). A Multilingual Prosodic Database. In *International Conference on Spoken Language Processing*, p. 3163–3166, Sydney, Australia.

Carrive J., Pachet F. & Ronfard R. (2000). CLAViS - A Temporal Reasoning System for Classification of Audiovisual Sequences. In *Content-Based Multimedia Information Access Conference (RIAO)*, College de France, Paris, France.

Dempster A. P., Laird N. M. & Rubin D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, **39 (Series B)**, 1–38.

Gauvain J. L., Lamel L. & Adda G. (1999). Systèmes de processus légers : concepts et exemples. In *International Workshop on Content-Based Multimedia Indexing*, p. 67–73, Toulouse, France : GDR-PRC ISIS.

Haidar S., Joly P. & Chebaro B. (2004). Detection Algorithm of Audiovisual Production Invariant. In *Workshop on Adaptative Multimedia Retrieval (AMR)*, p. 156–169, Valencia, Spain.

Houtgast T. & Steeneken J. M. (1985). A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria. *Journal of the Acoustical Society of America*, **77**(3), 1069–1077.

LLOYD S. P. (1957). Least Squares Quantization in PCM. Unpublished Bell Labs Technical Note (1957).

MODDEMEIJER R. (1989). On Estimation of Entropy and Mutual Information of Continuous Distributions. *Signal Processing*, **16**(3), 233–246.

PINQUIER J. & ANDRÉ-OBRECHT R. (2004). Jingle detection and identification in audio documents. In *International Conference on Audio, Speech and Signal Processing*, Montréal, Canada.

PINQUIER J., ROUAS J.-L. & ANDRÉ-OBRECHT R. (2003). Fusion de paramètres pour une classification automatique parole/musique robuste. *Technique et Science Informatiques (TSI)*, **22**(7-8), 831–852.

PINQUIER J., SÉNAC C. & ANDRÉ-OBRECHT R. (2002). Indexation de la bande sonore : recherche des composantes parole et musique. In *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, p. 163–170, Angers, France.

# EXPLOITATION DES RELATIONS TEMPORELLES ENTRE EVENEMENTS PRESENTS DANS LES DOCUMENTS AUDIOVISUELS

Zein Al Abidin IBRAHIM [1], Isabelle FERRANE [1] et Philippe JOLY [1]

[1] Université Paul Sabatier
IRIT, Toulouse
{Ibrahim, Ferrane, Joly}@irit.fr

**Résumé** : Le but de notre travail est de caractériser des structures intentionnelles dans des documents multimédia et particulièrement dans les vidéos. Pour cela, nous devons d'abord obtenir différentes segmentations du document pour étudier ensuite toutes les relations qui peuvent être observées entre ces segmentations. En ce qui nous concerne, nous ne disposons pas d'informations préalables concernant le type de la vidéo (sport, nouvelles…), la structure, ou le type de structure à rechercher. Notre travail est basé sur différents outils de segmentation fournissant automatiquement des unités temporelles en fonction de caractéristiques spécifiques essentiellement de bas niveau. A partir des segmentations obtenues, considérées deux à deux, nous effectuons une analyse des relations temporelles susceptibles d'exister entre les différents segments. Nous étudions la fréquence des réalisations de ces relations. Afin de donner une interprétation sémantique à ces observations, nous proposons une nouvelle représentation des relations temporelles, que nous appliquerons pour l'exemple aux relations de Allen. Après évocation du problème relatif aux erreurs de segmentation, nous présentons les premiers résultats obtenus après une première phase expérimentale produite sur des programmes TV et notamment des journaux télévisés.

**Mots-clés** : indexation multimédia, segmentation temporelle, relations de Allen.

## 1 Introduction

Beaucoup d'outils automatiques d'indexation de contenu audiovisuel produisent des résultats correspondant à des segments ou à des parties temporelles identifiant la présence ou non d'un objet spécifique. La plupart de ces outils se basent sur les résultats de l'extraction de caractéristiques bas niveaux : taux d'activité, couleur dominante, musique ou parole présents sur la bande sonore, textes ou caractères incrustés à l'écran … Leur but est de détecter des événements afin de créer des sommaires ou de déterminer des entrées temporelles dans un flot de données audiovisuelles. Ces premiers résultats sont ensuite utilisés soit pour constituer des index, soit pour servir de base à des étapes d'analyse ultérieures. La principale difficulté que l'on rencontre concerne l'interprétation sémantique que l'on peut attribuer aux index produits. Comme l'information pertinente utilisée durant le

processus d'analyse peut provenir de sources diverses (visuelles, audio, ou même textuelles) présentes dans un même document, on peut espérer que la combinaison de ces informations de bas-niveau puisse faire émerger une interprétation sémantique de plus haut niveau. De manière générale, les outils d'indexation peuvent être classés sur la base des caractéristiques utilisées dans le processus d'analyse. Dans la littérature, les caractéristiques extraites à partir de l'analyse de mouvement sont employées dans (Tovinkere & Qian, 2001) pour détecter un ensemble d'événements qui peuvent être présents dans une vidéo de matchs de football, et dans (Bonzanini et al., 2001) pour extraire un résumé et les moments importants de ce même type de vidéo. D'autres techniques sont basées sur des caractéristiques de couleur et de mouvement comme dans (Xie et al., 2002). Elles sont utilisées pour classifier une vidéo de match de football en deux phases : jeu et arrêt. La couleur et la forme sont employées dans (Avrithis et al., 2000) pour segmenter une vidéo de journaux télévisés en classes sémantiques. Une structuration des vidéos de matchs de basket-ball en classes sémantiques est proposée dans (Zhou et al., 2000) en se basant sur les règles du domaine et en combinant couleur, forme, mouvement et texture. D'autres techniques se basent sur des caractéristiques audio (Rui et al., 2000) et visent à résumer des vidéos de base-ball, ou à classifier les événements acoustiques rencontrés lors de matchs de football en classes sémantiques (Lefevre et al., 2002). Certains utilisent des caractéristiques multimodales (Eickeler & Muller, 1999) pour indexer les vidéos de journaux télévisés en détectant six classes sémantiques (Han et al., 2002) ou pour résumer automatiquement des vidéos de base-ball et des programmes de TV de Formule 1 (Petrovic et al., 2002).

Ces différentes techniques présentent plusieurs limitations notamment, l'utilisation de connaissances a priori sur le type d'analyse à effectuer et sur celui du document à traiter. Ceci suppose que l'on dispose en amont de l'indexation, d'une information précise sur ce que l'on va rechercher, sur le type de vidéo analysé ou sur les règles qui existent dans le domaine auquel appartient la vidéo (règles de jeu utilisées dans un match de football, règles utilisées dans la phase de production du type de document vidéo...). Dans ce cas, la portée de ces techniques est limitée à un contenu spécifique. Elles ne peuvent donc pas être employées telles quelles pour analyser de nouveaux types de contenu ni pour rechercher des événements qui ne sont pas prédéfinis. Des efforts de généralisation de ces techniques de détection d'événements ont été faits. Par exemple, dans (Duan et al., 2002), la généralisation a été faite à l'échelle d' « enregistrements d'événements sportifs » mais ceci reste malgré tout limité à un domaine spécifique (le domaine sportif).

Notre objectif est donc de se baser sur plusieurs systèmes de segmentation fournissant des informations sémantiques de plus ou moins bas niveau relatives à l'évolution du contenu d'un document, de les combiner afin d'observer les relations temporelles qui peuvent exister entre les événements ainsi identifiés. Ces observations permettront de déduire d'autres informations plus précises décrivant la structure temporelle du document. Notre approche est différente des techniques mentionnées ci-dessus car les structures que nous recherchons ne sont pas prédéfinies et aucune information

préalable n'est employée. Nous nous focalisons seulement sur l'étude des segments temporels que peuvent fournir différents systèmes de segmentation appliqués à différents médias (image, audio).

Cet article est organisé comme suit : dans la deuxième section, nous présentons différentes techniques de manipulation des relations temporelles et nous introduisons un mode de représentation graphique de celles-ci. Nous montrons ensuite comment cette représentation peut être employée pour identifier de manière générale des relations significatives entre événements, ce qui nous conduit à aborder les problèmes de quantification et de classification des relations. Enfin, à la fin de cette même section, nous illustrons notre approche en l'appliquant aux relations de Allen. Dans la troisième section, nous donnons quelques résultats des travaux expérimentaux menés jusqu'à présent sur un ensemble de quatre segmentations temporelles d'un journal télévisé. Enfin, dans la quatrième et dernière section, nous concluons et présentons nos travaux futurs dans ce domaine.

# 2 Information temporelle

## 2.1 Vue d'ensemble

L'analyse des relations temporelles entre les événements présents dans un même document audiovisuel est une question importante et ce pour différentes raisons. Les résultats d'une telle analyse peuvent être employés pour comparer le contenu d'un document donné à une structure temporelle prédéfinie (avec des HMM hiérarchiques par exemple) afin d'identifier des moments clés spécifiques, ou d'établir automatiquement une représentation temporelle de l'évolution du contenu. La situation actuelle démontre que de tels outils sont toujours construits sur la connaissance a priori de la façon dont les événements sont temporellement reliés entre eux dans les documents audiovisuels. Par exemple, nous pouvons employer le fait que dans un journal télévisé, la présence du présentateur alterne avec les reportages, ou que dans une émission de variété, la performance d'un artiste est suivie d'applaudissements présents sur la bande sonore.
Pour étendre l'analyse temporelle du contenu audio ou visuel d'un document aux relations susceptibles d'exister entre n'importe quel genre d'événements, qu'elles soient flagrantes ou subtiles et imprévisibles, il est nécessaire de disposer d'outils de représentation et de raisonnement temporels.

Hayes a introduit six notions différentes pour représenter des relations temporelles (Hayes, 1995) à savoir : la dimension physique de base, la « time-line », les intervalles de temps, les points temporels, la quantité de temps ou durée et les positions de temps. Ce problème a également été abordé par plusieurs chercheurs. Nous pouvons trouver un état de l'art des différentes approches de représentation et de raisonnement

temporels en se référant à Chittaro et Montanari ((Chittaro & Montanari, 1996) et (Chittaro & Montanari, 2000)), Vila (Vila, 1994), et Pani (Pani, 2001).

Les modèles existants et permettant d'exprimer les relations temporelles peuvent être divisés en deux catégories : les modèles basés sur la notion de point temporel (Vilain & Kautz, 1986) et ceux basés sur la notion d'intervalle (Allen, 1983).

Dans le premier type de modèles, les points sont des unités élémentaires répartis le long de l'axe du temps. Chaque événement est associé à un point temporel. Soient deux événements **e1** et **e2**, trois relations temporelles peuvent être déterminées entre eux. Un événement peut être *avant* (<), *après* (>) ou *simultané* (=) à un deuxième événement. Ces relations sont des relations basées sur la notion de point temporel. Un exemple de représentation de ce type est la « time-line », sur laquelle les objets sont placés sur plusieurs axes de temps. Cette représentation a par la suite été également employée comme représentation basée sur la notion d'intervalle. Nous pouvons trouver le modèle de « time-line » utilisé dans diverses applications telles que HyTime (HyTime, 1992).

Les modèles basés sur la notion d'intervalle considèrent les entités élémentaires comme des intervalles de temps qui peuvent être ordonnés selon différentes relations. Les modèles existants sont principalement basés sur les relations définies par Allen dans (Allen, 1983).

## 2.2 Représentation et raisonnement temporels

Considérons deux segmentations temporelles d'un même document vidéo, **S1** et **S2**, effectuées pour procéder à l'analyse du document. Cette première étape produit des segments qui peuvent être vus comme des intervalles temporels localisant chacun une partie de la vidéo dans laquelle un événement spécifique a été détecté. Chaque système de segmentation produit une séquence d'intervalles où un seul type d'événement se produit (effets de transition progressifs, présence d'un personnage à l'image, présence de musique sur la bande sonore, etc..). Ainsi, on peut considérer le résultat d'une segmentation donnée, comme un ensemble de segments temporellement disjoints. Les segmentations S1 et S2 comportant respectivement N et M segments successifs seront définies par : $S1 = \{s_{1i}\}_{i\in[1,N]}$ et $S2 = \{s_{2j}\}_{j\in[1,M]}$

Un intervalle temporel est caractérisé par deux points correspondant à ses extrémités. Soit $s_{1i}$ et $s_{2j}$ deux segments issus respectivement de la segmentation S1 et de la segmentation S2. Chacun de ces segments est caractérisé par son point de début (d) et son point de fin (f), soit respectivement $[s_{1id}, s_{1if}]$ et $[s_{2jd}, s_{2jf}]$. Nous pouvons représenter la relation temporelle entre ces segments à l'aide de trois variables, comme proposé dans (Moulin, 1992) :

1) **Lap** = $s_{2jd} - s_{1if}$
2) **DB** = $s_{1id} - s_{2jd}$
3) **DE** = $s_{2jf} - s_{1if}$

Ainsi, une relation temporelle entre deux segments peut être représentée dans un espace à trois dimensions. Un point dans cet espace détermine une relation entre deux intervalles. Pour deux segmentations **S1** et **S2**, les trois paramètres caractérisant chaque couple de segments ($s_{1i}$, $s_{2j}$) peuvent être évalués et représentés par un point dans l'espace 3D.
A chaque point de l'espace 3D (i.e. pour chaque relation temporelle potentielle entre deux segments), nous associons un accumulateur qui comptabilise les votes c'est-à-dire qui compte le nombre de fois où la relation associée est observée entre deux segmentations. Ces accumulateurs sont regroupés en une matrice appelée par la suite Matrice des Relations Temporelles (MRT). Cette matrice peut être employée directement pour déterminer les fréquences des relations potentielles ainsi que pour observer les distributions des votes. La mise en évidence de distributions remarquables, permettra d'identifier une règle générale caractérisant le comportement temporel des événements segmentés.
La taille de la MRT est directement liée à la durée du document et peut par conséquent être très grande. Ainsi, le premier problème à surmonter est la quantification de l'espace 3D afin de produire une matrice de taille acceptable. Une fois que la matrice est créée, initialisée et remplie (avec les différents votes), l'étape de quantification peut être exécutée.

### 2.2.1 Quantification

L'étape de quantification de la matrice dépend de l'échelle des caractéristiques de bas niveau. En ce qui concerne le contenu visuel d'un document, l'extraction de caractéristiques de bas niveau peut associer une valeur à chaque image ou à des fenêtres d'une durée de 1 seconde. Dans le cas de l'analyse audio, les résultats produits ne correspondent pas nécessairement aux mêmes unités de temps, la durée des segments sur lesquels les caractéristiques sont déterminées étant relativement variable. Par conséquent, nous devons dans chaque cas, définir des intervalles basés sur la plus grande échelle temporelle employée pour exprimer les caractéristiques.
D'une façon plus générale, la quantification de cet espace mène aux mêmes problèmes que ceux généralement identifiés dans des méthodes de vote. En particulier, la taille de la matrice doit être limitée. Comme la variation maximale des paramètres est inférieure ou égale à la différence entre le début du premier intervalle et la fin du second nous pouvons a priori identifier les frontières de cet espace. En outre, dans le cas d'un espace de dimension élevée, nous pouvons directement appliquer un processus hiérarchique de discrétisation au lieu de procéder à une quantification brute et ainsi se concentrer progressivement sur les sous parties de l'espace qui reçoivent plus de voix que les autres (Li & Lavin, 1986).

### 2.2.2 Classification

Une fois que l'étape de vote a été exécutée, c'est-à-dire lorsque tous les couples possibles de segments ont été traités et que les votes ont été effectués, la MRT doit être analysée pour identifier par exemple les relations les plus fréquentes entre les

caractéristiques prises en compte. À la différence d'autres techniques de vote, s'intéresser uniquement à la valeur maximale n'est pas suffisant pour identifier entièrement une relation. En effet, la plupart des relations sémantiques temporelles déterminent des sous espaces de la MRT dans lesquels les votes sont distribués. Ainsi, la première étape de l'analyse de la MRT est de localiser ces zones. Cette localisation peut être réalisée par des méthodes de clustering ou des outils de séparation de classes.

Une autre approche consiste à définir a priori les relations sémantiques à observer, comme par exemple en prenant les relations de Allen. Cette approche consiste alors à identifier les sous-parties disjointes de l'espace de vote qui peuvent être associées aux relations remarquables comme cela est illustré dans l'exemple donné en section 2.2.3.

Ensuite, le nombre d'occurrence de chaque relation **R** observée entre deux caractéristiques est calculé en effectuant la somme des votes contenus dans la sous partie associée.

### 2.2.3 Exemple

Allen a proposé un ensemble complet de relations temporelles pouvant exister entre deux intervalles. Ainsi, pour deux intervalles donnés, il définit treize possibilités distinctes de relier temporellement ces segments. Dans le tableau 1 les douze premières lignes représentent les six relations directes et leur relation inverse respective et la dernière ligne correspond au cas où les deux segments débutent et se terminent en même temps. Puisqu'un intervalle est défini par deux points (son début et sa fin) on peut ramener un modèle basé sur des relations entre intervalles à un modèle basé sur les points en considérant une relation entre intervalles comme une conjonction de relations entre les points correspondants aux extrémités des segments observés ($[s_{1id}, s_{1if}]$ et $[s_{2jd}, s_{2jf}]$ cf . Tableau 1).
Dans la relation '*before* (<) ainsi que pour son inverse, (>) nous ajoutons une contrainte, nommée 'α', pour limiter la détection de ces deux types de relation à deux intervalles dont la comparaison reste significative et porteuse d'informations pertinentes à propos de la structure du contenu.

| **Relation** | **Symbole et *Inverse*** | ***Notation Point*** | **Exemple** |
|---|---|---|---|
| $s_{1i}$ **before (α)** $s_{2j}$ | **<**<br>**>** | $s_{1id}<s_{1if}<s_{2jd}<s_{2jf}$<br>**&** $(0<s_{2jd} - s_{1if} \leq \alpha)$ | **AAA** $\Leftrightarrow$ **BBBB**<br>$d = s_{2jd} - s_{1if} \leq \alpha$ |
| $s_{1i}$ **meets** $s_{2j}$ | **m**<br>***mi*** | $s_{1id}<s_{1if}=s_{2jd}<s_{2jf}$ | **AAAAA**<br>**BBBB** |
| $s_{1i}$ **overlaps** $s_{2j}$ | **o**<br>***oi*** | $s_{1id}<s_{2jd}<s_{1if}<s_{2jf}$ | **AAAAA**<br>**BBBBB** |
| $s_{1i}$ **starts** $s_{2j}$ | **s**<br>***si*** | $s_{1id}=s_{2jd}<s_{1if}<s_{2jf}$ | **AAAA**<br>**BBBBBBBB** |
| $s_{1i}$ **finishes** $s_{2j}$ | **f**<br>***fi*** | $s_{2jd}<s_{1id}<s_{1if}=s_{2jf}$ | **AAA**<br>**BBBBBBB** |

| | | | |
|---|---|---|---|
| $s_{1i}$ **equals** $s_{2j}$ | **=**<br>**=** | $s_{1id}=s_{2jd}<s_{1if}=s_{2jf}$ | **AAAAAA**<br>**BBBBBB** |
| $s_{1i}$ **during** $s_{2j}$ | **d**<br>***di*** | $s_{2jd}<s_{1id}<s_{1if}<s_{2jf}$ | **AAAAA**<br>**BBBBBBBBB** |

Tableau 1

Si nous représentons chaque relation temporelle entre deux intervalles à l'aide des trois paramètres **Lap**, **DE**, **DB** définis plus haut, et que nous appliquons cette représentation aux relations de Allen, nous constatons que celles-ci définissent des contraintes entre ces paramètres comme nous le montrons dans le Tableau 2.

Par exemple, la relation '*during'* correspond à la définition suivante:

$$\mathbf{s_{2jd} < s_{1id} < s_{1if} < s_{2jf}}$$

D'où on peut déduire que :

$s_{2jd} - s_{1if} < s_{1id} - s_{1if} < s_{1if} - s_{1if} < s_{2jf} - s_{1if}$ => **Lap < 0 < DE**.

$s_{2jd} - s_{2jd} < s_{1id} - s_{2jd} < s_{1if} - s_{2jd} < s_{2jf} - s_{2jd}$ => 0 < **DB < -Lap**

Pour la relation *'meet inverse'* (mi), nous avons les contraintes suivantes:

$$\mathbf{s_{2jd} < s_{2jf} = s_{1id} < s_{1if}}$$

$s_{2jd} - s_{2jd} < s_{1id} - s_{2jd} = s_{2jf} - s_{2jd} < s_{1if} - s_{2jd}$ => 0 < **DB**.

$s_{2jd} - s_{1if} < s_{2jf} - s_{1if} < s_{1if} - s_{1if}$ => **DE** < 0.

$s_{2jd} - s_{1if} < s_{2jd} - s_{1if} = s_{1id} - s_{1if} < 0$ => **Lap** < 0

**DE – DB** = $s_{2jf} - s_{1if} - s_{1id} + s_{2jd} = (s_{2jf} - s_{1id}) + (s_{2jd} - s_{1if})$ = **Lap**,
$s_{2jf} - s_{1id} = 0$ parce que $s_{2j}$ 'meet' $s_{1i}$.

De la même manière, nous pouvons déduire des contraintes similaires pour chacune des autres relations de Allen. Ces contraintes définissent des sous-espaces dans la MRT localisant les votes de chaque relation.

| ***Relation*** | ***Lap*** | ***DB*** | ***DE*** |
|---|---|---|---|
| **<** | **0 < Lap ≤ α** | **DB< -Lap** | **DE > Lap** |
| ***m*** | **Lap = 0** | **DB < 0** | **DE > 0** |
| ***o*** | **Lap < 0** | **DB < 0** | **DE > 0** |
| ***s*** | **Lap < 0** | **DB = 0** | **DE > 0** |
| ***f*** | **Lap < 0** | **DB > 0** | **DE = 0** |
| **=** | **Lap < 0** | **DB = 0** | **DE = 0** |
| ***d*** | **Lap < 0** | **DB < -Lap** | **DE > 0** |
| **>** | **DE-DB < Lap < 0 et 0<DB –DE+Lap ≤ α** | **DB > 0** | **DE < 0** |
| ***mi*** | **Lap =DE-DB** | **DB > 0** | **DE < 0** |
| ***oi*** | **Lap < DE-DB <0** | **DB > 0** | **DE < 0** |
| ***si*** | **Lap < DE** | **DB = 0** | **DE < 0** |

| | | | |
|---|---|---|---|
| ***fi*** | **Lap < 0** | **DB < 0** | **DE = 0** |
| ***di*** | **Lap < DE** | **DB < 0** | **DE < 0** |

Tableau 2 : Contraintes sur les paramètres caractérisant les relations de Allen

Une quantification a priori de l'espace utilisant les relations de Allen est définie par les règles données dans le Tableau 2.

Pour simplifier la représentation dans l'espace 3D, nous établissons la correspondance suivante : DE=x, DB=y, et Lap = z ;

Considérons la région définie par la relation '*before'*. Les contraintes définies dans le tableau 2 indiquent qu'une relation établie entre deux intervalles $s_{1i}$ et $s_{2j}$ sera la relation '*before'* si et seulement si :

1) $0 < z \leq \alpha$
2) $y < -z$
3) $x > z$

Ainsi, les votes qui correspondent à la relation '*before'* seront cumulés dans la zone délimitée par les contraintes mentionnées ci-dessus et représentée sur la figure 1.

La relation '*meet*' (Fig. 2) correspond à une zone restreinte au plan défini par l'équation: z = 0. Plus précisément, la relation '*meet'* peut être établie entre $s_{1i}$ et $s_{2j}$ si et seulement si on vérifie :

1) $z = 0$
2) $y < 0$
3) $x > 0$

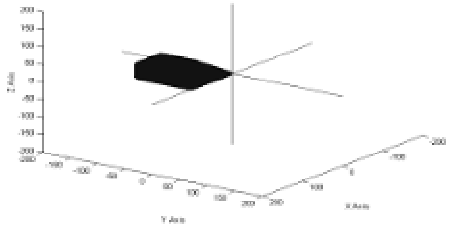


Fig. 1 - La relation *before*

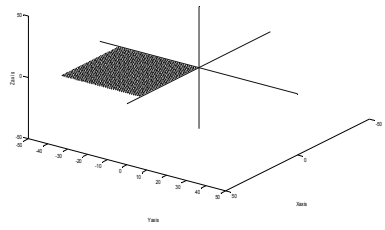


Fig. 2 - La relation *meet*

On peut également considérer que la relation '*equal'* peut être établie entre $s_{1i}$ et $s_{2j}$ ssi on vérifie :

1) $z < 0$
2) $y = 0$
3) $x = 0$

ce qui correspond à la demi-droite représentée dans la figure 3.

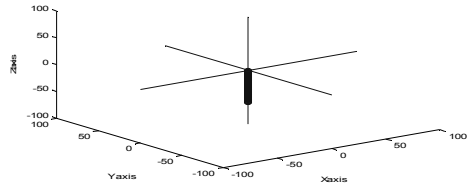



Fig. 3 - La relation *equal*

Suivant le même principe, nous pouvons identifier les régions de la MRT correspondant à toutes les autres relations.

Avec cette représentation, nous pouvons ainsi établir un lien entre une observation donnée entre deux intervalles et une relation. En d'autres termes, les paramètres définissant l'observation d'une relation et vérifiant les équations définissant une zone donnée permettront d'identifier la relation sémantique correspondante.

## 2.3 Gestion d'erreur

Une segmentation temporelle, si elle est obtenue automatiquement, peut être imprécise. Plusieurs sources d'imprécision existent. Des erreurs peuvent être dues aux performances des outils d'extraction des caractéristiques de bas niveau utilisés pour traiter le flot de données audiovisuelles. Ceci peut conduire à voter pour une relation qui en réalité n'est pas la relation qui devrait être observée mais une relation voisine. L'effet de bord ainsi introduit par l'imprécision de certains outils utilisés pour la segmentation, doit être pris en compte. On ne peut pas être sûr à cent pour cent de la légitimité du vote enregistré. Une approche possible pour prendre en compte cette incertitude consisterait à considérer une distribution floue des votes sur les relations voisines.
Dans le cas des relations de Allen et comme pour la plupart des approches dans ce domaine, nous pouvons employer le principe des relations voisines de Freska (Freska, 1992). Soient A et B, deux événements vérifiant la relation temporelle '*meet'*. En déplaçant ou en déformant légèrement les segments, nous pouvons changer la relation en *'before'* ou *'overlap'*. Par conséquent, les relations *'before'* et *'overlap'* sont considérées comme relations conceptuellement voisines de la relation *'meet'*. Au contraire, la relation *'equal'*, par exemple, n'est pas une voisine conceptuelle de *`meet'* , car elle ne peut pas être obtenue directement à partir *de 'meet'* par

déformation ou translation temporelle des intervalles. Dans notre cas, chaque relation est représentée par une zone et ses voisines sont les zones adjacentes.

Une extension topologique du principe de voisinage défini par Freska à n'importe quelle relation peut ici être formellement définie par :

Soit **Ri** (Xi, Yi, Zi) et **Rj** (Xj, Yj, Zj) deux zones compactes et soit $\mathbf{R_k}$ = Ri ∩ Rj
ou $X_k$ = (Xi ∩ Xj), $Y_k$ = (Yi ∩ Yj), et $Z_k$ = (Zi ∩ Zj).
**Ri** a **Rj** comme voisine directe si un seul des paramètres de **Rk** est vide.

Par exemple, si Ri = ***meet*** . alors Xi correspond à **DE** > 0, Yi à **DB** < 0, et Zi à **Lap** = 0.

Ri (Xi, Yi, Zi) = **meet (** ]0 +∞[ **,** ]-∞ 0[ , {0}**)**

Si Rj = ***overlap*** alors Xj correspond à **DE** > 0, Yj à **DB** < 0, et Zj à **Lap** < 0.
Rj (Xj, Yj, Zj) = **O (** ]0 +∞[ , ]-∞ 0[ , ]-∞ 0[ **)**

Nous avons alors (Xi ∩ Xj) = ]0 +∞[ , (Yi ∩ Yj) = ]-∞ 0[, et (Zi ∩ Zj) = Ø. Comme on a seulement (Zi ∩ Zj) = Ø, alors ***meet*** et ***overlap*** sont des voisines directes.

Ce n'est pas le cas pour les relations ***Meet*** et ***During*** pour lesquelles deux paramètres sont vide : $Y_k$ = Ø, et $Z_k$ = Ø.

Ces liens de voisinages entre relations sont illustrés dans la figure 4 où sont représentées les zones correspondant aux relations 'meet' (gris intermédiaire), 'overlap' (gris clair), et 'start' (gris foncé).

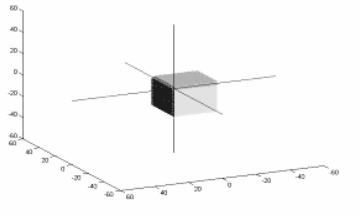


Fig. 4 : liens de voisinage entre relations Meet, Overlap et start

## 3 Expérimentation

Nous avons calculé plusieurs matrices de type MRT sur des segmentations temporelles effectuées sur des vidéos de journaux télévisés. Cinq segmentations différentes ont été réalisées sur un même document. Elles sont relatives à la présence

ou non du présentateur à l'écran et, du point de vue sonore, à la présence ou non de parole, de silence, de musique et d'applaudissements. Nous avons construit quatre MRT pour observer les relations temporelles entre la segmentation visuelle (présentateur) et chacune des segmentations sonores (parole, silence, musique, applaudissements). A la différence de l'exemple donné dans la section 3, nous n'avons pas de connaissance a priori sur les relations potentiellement observables. Seule l'étude des résultats obtenus par le calcul des MRT doit nous permettre d'observer des régularités et d'en déduire la présence de relations pertinentes entre segments.

En observant la première MRT dont la représentation est données figure 5, nous constatons que les points représentant les relations entre le présentateur et les segments de la parole sont distribués le long de lignes parallèles, toutes inclues dans le plan d'équation $z = ay + b$ ; pour x arbitraire. Les cercles représentent les points exclus après l'étape de quantification et donc considérés comme non pertinents (i.e. quand $Lap > \alpha$). Cette distribution peut être également observée dans la figure 9.

Après la projection de cette MRT sur l'axe XOZ (Fig.6), nous observons également que la plupart des points se retrouvent sur une même ligne passant par le centre de l'axe XOZ.

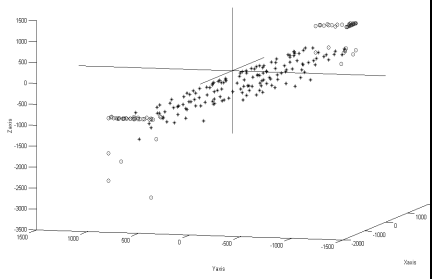


Fig. 5 : Relation entre présentateur et parole

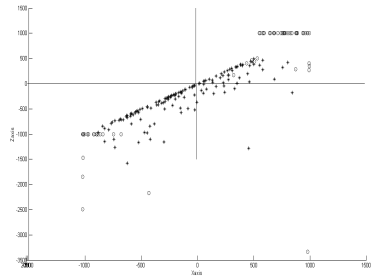


Fig. 6 : Projection sur l'axe XOZ des points de la figure 5

Dans la figure 7 nous avons représenté les relations qui ont pu être observées entre le présentateur et les segments musicaux. Dans la figure 8 sont représentées celles observées entre le présentateur et les applaudissements. Nous pouvons remarquer que des relations entre les événements « présence du présentateur » et « segments musicaux » existent, ce qui n'est quasiment pas le cas pour les relations entre « présentateur » et « applaudissements ». Les applaudissements détectés sont en fait tous des erreurs occasionnelles de l'outil de segmentation concerné. Bien que lorsque le présentateur parle, la présence de musique soit très rare, il arrive que le journal soit

entrecoupé de publicités souvent accompagnées de séquences musicales (jingles par exemple). En ce qui concerne la relation entre « présentateur » et « silence » (fig. 9), le nombre de points augmente et la projection des points sur l'axe XOZ (fig. 10) est également une droite qui passe par le centre. Une transformation de Hough pourrait être employée pour identifier cette droite ou celles qui apparaissent sur la figure 7. Ce type d'analyse fournit des informations plus précises sur le contenu puisque nous avons des relations distribuées suivant des types de zones larges et clairement identifiables (droites, plans, …) qui ne correspondent pas aux régions identifiées par les relations de Allen.

Dans la figure 10, nous pouvons observer que les points projetés sont distribués sur le plan où x ≈ z. Cela signifie qu'effectivement, DE ≈ LAP, et donc que $s_{2jd} - s_{1if} \approx s_{2jd} - s_{1if}$. Nous en déduisons que $s_{2jf} \approx s_{2jd}$ ce qui signifie que les segments de la seconde caractéristique sont très courts. Celle-ci correspond aux segments de silence présents dans un journal télévisé, qui sont effectivement très courts dans ce type de document.

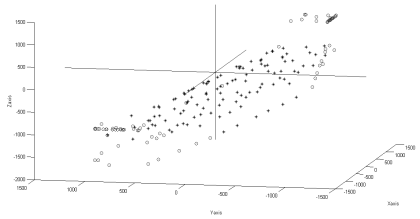


Fig. 7 : Relations entre présentateur et musique

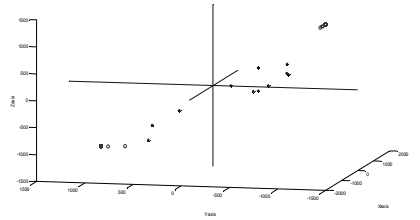


Fig. 8 : Relations entre présentateur et applaudissement

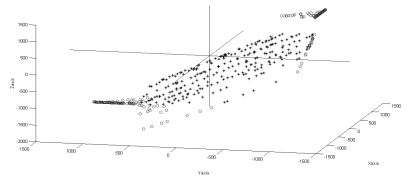


Fig. 9 : Relations entre présentateur et silence

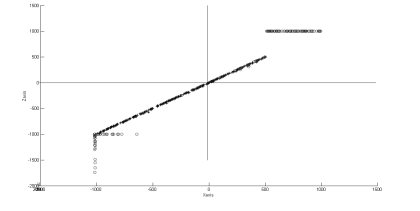


Fig.10 : Projection sur l'axe XOZ des points de la figure 9

## 4 Conclusion et travaux futurs

Nous avons présentés dans cet article une nouvelle technique pour mettre en évidence les relations temporelles significatives entre des événements issus de la segmentation

produite par des outils automatiques. Les relations temporelles entre deux intervalles appartenant à deux segmentations différentes d'un même document sont représentées par des points dans un espace à trois dimensions. L'espace des observations peut être discrétisé en différentes zones, chacune d'elles pouvant représenter une relation sémantique. Cette discrétisation peut être effectuée en ayant recours à une méthode de classification traditionnelle ou bien en partant de relations temporelles déjà connues comme nous l'avons fait à titre d'exemple en utilisant les relations temporelles de Allen. Nous avons ensuite présenté les premiers résultats d'une expérimentation effectuée sur des documents vidéos, notamment des journaux télévisés, ce qui nous a permis d'observer la distribution des points dans l'espace de représentation.

Plusieurs perspectives peuvent être envisagées pour la poursuite de ce travail. Un de nos objectifs est d'aborder le problème lié aux performances plus ou moins bonnes des outils de segmentation utilisés, ce qui peut influer sur la fiabilité des observations réalisées. Comme mentionné précédemment, les votes enregistrés dans la MRT peuvent être distribués dans le voisinage d'une relation, la notion de voisinage étant définie par exemple suivant le graphe de voisinage de Freska appliqué aux relations de Allen. La détermination des liens de voisinage entre relations dans l'espace de la MRT peut être réalisé en utilisant la distance entre les zones de l'espace à trois dimensions les plus proches. Un arbre de voisinage peut être automatiquement construit reliant chaque sous-espace de la MRT. Les poids associés à un vote peuvent être distribués d'une manière plus précise tout en utilisant la topologie de la MRT au lieu d'employer un arbre de voisinage. Le calcul des poids peuvent dépendre de différents critères : de la position du point dans une zone (c.-à-d. de la distance entre ce point et le centre de la zone), de la distance qui la sépare à d'autres zones, de la taille des zones, du nombre de points inclus…

Nous avons également l'intention d'explorer la conjonction des relations observées d'une manière hiérarchique. Pouvoir manipuler une conjonction d'un grand ensemble de relations devrait nous permettre d'identifier des événements temporels complexes.

Enfin, notre objectif à plus long terme est d'explorer l'utilisation de la MRT comme modèle qui puisse caractériser l'évolution temporelle de différents types de documents.

# Références

Allen J. F. (1983). Maintaining Knowledge about Temporal Intervals (Tome 26 (11)). Communication of the ACM. p. 832 – 843.

Avrithis Y., Tsapatsoulis N. & Kollias S.(2000). Broadcast News Parsing Using Visual Cues: A Robust Face Detection Approach. In IEEE International Conference on Multimedia and Expo. New York City, USA.

Bonzanini A., Leonardi R., & Migliorati P. (2001). Exploitation of Temporal Dependencies of Descriptors to Extract Semantic information. International Workshop on Very Low Bitrate Video Coding. Athen, Greece.

Chittaro L. & Montanari A. (1996). Trends in Temporal Representation and Reasoning (Tome 11(3)). The Knowledge Engineering Review. p. 281-288.

Chittaro L. & Montanari A. (2000). Temporal Representation and Reasoning in Artificial Intelligence: Issues and Approaches (Tome 28). Annals of Mathematics and Artificial Intelligence. p. 47-106.

Duan L., Xu M., Xiao-Dong Yu, & Qi Tian (2002). A unified framework for semantic shot classification in sports videos. In Proc. of the tenth ACM international conference on Multimedia. p. 219-220. Juan-les-Pins, France.

Eickeler S. & Muller S. (1999). Content-Based Video Indexing of TV Broadcast News Using Hidden Markov Models (Tome 6). In Proc. IEEE ICASSP. P. 2997-3000. Phoenix, USA.

Freska C. (1992). Temporal Reasoning Based on Semi-intervals (Tome 54). Artificial Intelligence. p.199-227.

Han M., Hua W., Xu W., & Gong Y. (2002). An integrated baseball digest system using maximum entropy method. In Proc. ACM Multimedia 2002. p. 347-350. Juan Les Pins, France.

Hayes J. Patrick. (1995). A Catalog of temporal theories. Technical report UIUC-BI-AI- 96-01, University of Illinois.

HyTime (1992) Information Technology, "Hypermedia / Time-based Structuring Language (HyTime)", ISO/IEC 10743.

Li H. & Lavin M. A. (1986). Fast Hough Transform: A Hierarchical Approach (Tome 36). Journal on Graphical Models and Image Processing (CVGIP). p.139-161.

Lefevre S., Maillard B., & Vincent N. (2002). 3 classes segmentation for analysis of football audio sequences. In Proc. ICDSP'2002. Santorin, Greece.

Moulin B. (1992). Conceptual graph approach for the representation of temporal information in discourse (Tome 5 (3)). Knowledge based systems. p 183 –192.

Pani A. K. (2001). Temporal representation and reasoning in artificial intelligence: A review. Mathematical and Computer Modelling. p. 55–80.

Petrovic M., Mihajlovic V., Jonker W., & Djordievic-Kajan S. (2002). Multi-modal extraction of highlights from tv formula 1 programs. In Proc. ICME'2002. Lausanne, Switzerland.

Rui Y., Gupta A., & Acero A. (2002). Automatically extracting highlights for TV baseball programs. In Proc. of the eight ACM international conf. on Mult. p. 105–115. California, USA.

Tovinkere V., Qian R. J. (2001). Detecting Semantic Events in Soccer Games: Toward a Complete Solution. In Proc. ICME'2001. p. 1040-1043. Tokyo, Japan.

Vila L. (1994). A Survey on Temporal Reasoning in Artificial Intelligence (Tome 7(1)). Artificial Intelligence Communications. p. 4 -28.

Vilain M., & Kautz H. A. (1986). Constraint propagation algorithms for temporal reasoning. In AAAI-86. p. 132-144.

Xie L., Chang S-F., Divakaran A., and Sun H.(2002). Structure analysis of soccer video with Hidden Markov Models. In Proc. International Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Orlando, USA.

Zhou W., Vellaikal A., & Kuo C.-C. J.(2000). Rule-based Video Classification System for Basketball Video Indexing. In Proc. of the 2000 ACM Mult. Workshops. p. 213-216. Los angelos, USA.

# Ontologies et description du contenu de documents AV : une expérimentation dans le domaine médical*

Antoine Isaac[1,2] et Raphaël Troncy[3]

[1] Institut National de l'Audiovisuel, Direction de la Recherche
4, Av. de l'Europe - 94366 Bry-sur-Marne, France
aisaac@ina.fr

[2] Université de Paris-Sorbonne, LaLICC

[3] ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
raphael.troncy@isti.cnr.it
http://nmis.isti.cnr.it/troncy/

**Résumé** : L'utilisation de connaissances formalisées dans le processus de description de contenus audiovisuels est une technique prometteuse qui fournit déjà des résultats encourageants. Dans cet article, nous présentons une expérimentation concernant la description de documents audiovisuels ayant pour thème la médecine. Cette description du contenu d'émissions télévisées repose sur la notion de patrons d'indexation s'appuyant sur des scénarios d'utilisation existants, et exploite les technologies issues du web sémantique. Nous montrons que la combinaison de plusieurs ontologies et de règles d'inférence permet une description structurée et une recherche des séquences audiovisuelles plus complète, la base d'index pouvant être augmentée de connaissances additionnelles selon l'expressivité du langage retenu et les spécifications contenues dans les ontologies.

## 1 Introduction : mise en place de l'expérimentation

Alors que les documents audiovisuels sont de plus en plus accessibles, notamment *via* le web, leur traitement pose toujours problème. En particulier, rechercher des séquences pertinentes en suivant des critères liés au contenu n'est pas trivial, ce qui peut nuire par exemple à la réutilisation d'éléments préexistants dans une nouvelle production. Dans cette optique, le traitement du contenu

*Ce travail a été partiellement soutenu et financé dans le cadre d'une bourse ERCIM.

audiovisuel à l'aide de techniques à base de connaissances constitue une piste de recherche séduisante. La communauté du *web sémantique* a récemment proposé des standards permettant de représenter des ontologies (OWL, 2004) et des annotations (RDF, 2004), et a produit des outils pour formuler des requêtes et conduire des raisonnements sur des bases de connaissances. Nous avons déjà montré comment l'accès au contenu multimédia peut bénéficier de ces techniques, tout en mettant l'accent sur les problèmes importants à résoudre dans ce cadre (Troncy, 2003; Isaac *et al.*, 2004; Troncy, 2004). Dans cet article, nous allons présenter une expérimentation basée sur des scénarios d'utilisation existants, où ces techniques sont employées pour décrire en détails le contenu d'un corpus de programmes TV. Nous montrons que la combinaison de plusieurs ontologies et de règles d'inférence améliore la description et la recherche des séquences concernées.

Notre expérimentation se concentre sur des documentaires liés à la médecine. Des fonds de l'INA, nous avons extrait une trentaine de documents AV, en majorité des magazines, d'à peu près une heure chacun. Nous en avons ensuite sélectionné la moitié liée aux thèmes du cœur ou de la chirurgie cardiaque. Il s'agit donc d'une collection plutôt homogène, ce qui facilite la recherche d'une ontologie thématique adaptée. De fait, comme le sujet de la médecine a attiré l'attention d'un grand nombre de chercheurs ces dernières années, de nombreuses ressources ontologiques sont disponibles dans ce domaine. De plus, comme ces programmes devaient être diffusés *via* la télévision, ce sont aussi de bons exemples sur la manière dont les procédés audiovisuels sont utilisés pour vulgariser des sujets scientifiques complexes.

Les applications qui utilisent les documents AV (insertions d'extraits AV dans de nouveaux documents, utilisation de l'audiovisuel comme support pédagogique ou comme matériau de recherche...) peuvent être intéressées par différents aspects. Chacune apporte son point de vue, et est uniquement concernée par les éléments d'information qui correspondent à ses besoins. Un institut comme l'INA doit collecter et décrire un patrimoine audiovisuel. Il est ainsi concerné à la fois par la forme et le contenu des documents, et insiste sur un point de vue plutôt archivistique. Par conséquent, les descriptions doivent mélanger des éléments strictement orientés "description audiovisuelle" (par exemple l'affirmation qu'un document contient des séquences particulières comme des interviews et utilise des techniques comme l'animation) et des notions relevant d'un domaine donné. Ainsi, savoir qu'une séquence inclut plusieurs gros plans d'une opération chirurgicale rend celle-ci intéressante en tant que matériel pédagogique scientifique. Cette séquence, décrite correctement, pourrait alors être retrouvée et ré-utilisée dans un autre document. Notre expérimentation suit des principes de description observés dans le cadre de l'activité documentaire à l'INA et du projet Opales[1] (Isaac *et al.*, 2004), ce qui en fait un scénario réaliste de la manière dont les techniques du web sémantique et les pratiques traditionnelles peuvent être utilisées de concert.

Dans la section suivante, nous détaillons les ressources ontologiques nécessaires

[1] `http://opales.ina.fr/public/`

à la description de notre corpus. Dans la section 3, nous montrons comment ces ontologies sont utilisées pour concevoir des indexations ciblant à la fois la structure et le contenu des documents. Dans la section 4, nous discutons des types de raisonnement que nous avons pu mettre en œuvre en fonction du langage de représentation d'ontologies et de l'utilisation de règles d'inférences additionnelles. Dans la section 5, nous présentons quelques travaux existants dans le champ de la description de contenu multimédia utilisant des connaissances, avant de conclure en section 6.

## 2 Ressources ontologiques

Nous avons déjà montré dans (Troncy, 2003, 2004) l'intérêt d'articuler une ontologie dédiée à l'audiovisuel avec des ontologies thématiques dans le but de produire des descriptions qui correspondent vraiment à des applications documentaires spécifiques. Cette méthode permet de rentabiliser notre expérience en description AV : un tel savoir, modulaire, peut être aisément adapté d'une application à une autre. Nous avons donc proposé dans (Isaac & Troncy, 2004) une ontologie *noyau* pour la description AV, utilisable pour une large gamme d'applications.

Cette ontologie se concentre sur la caractérisation d'éléments documentaires : le concept principal est celui d'*objet de production AV*, qui représente la notion même de document AV. La première distinction intervient entre *programmes* (entités plutôt indépendantes du point de vue de la production et de la diffusion) et *séquences* (parties de programmes ou d'autres séquences). Ces concepts sont ensuite spécialisés en fonction de traits différentiels liés à la forme ou au contenu, pour obtenir le schéma classificatoire commun à tous les usages : les *genres*. Par exemple, les programmes sont répartis entre *composites* et *simples*, les premiers, contrairement aux seconds, étant composés d'une suite d'éléments autonomes du point de vue de la forme et du contenu. Ils sont ensuite classés suivant leur longueur et leur contenu général (fiction, information, divertissement). Après quelques étapes additionnelles de spécialisation, on peut trouver les genres télévisuels courants : *comédie de situation*, *documentaire*, *spectacle TV*...

L'ontologie introduit également les notions utilisées pour préciser les caractéristiques des objets AV. Tout d'abord, nous avons introduit une hiérarchie des rôles que les personnes peuvent jouer dans un programme, soit en tant qu'auteurs (*producteur*, *réalisateur*) mentionnés à cause de leur importance dans la production du programme, soit comme participants (*animateur*, *acteur*), apparaissant dans la description parce que visibles dans le document. Ensuite, nous pouvons trouver un ensemble important de propriétés qui reflètent diverses préoccupations ou modalités de la production (filmage, comme pour *mouvement de caméra* ; montage ou post-production, comme pour *insertion de texte*) et la diffusion (*date de diffusion*, *périodicité*, *public visé*...). Une typologie des thèmes généraux que peut aborder un document vient compléter l'ontologie.

Nous attirons l'attention sur le fait que cette ontologie a été construite en accord avec un certain nombre de principes méthodologiques, expliqués dans

(Bachimont *et al.*, 2002; Isaac & Troncy, 2004). Les concepts sont en particulier rattachés à des patrons de conception ontologiques de haut niveau, pour augmenter le potentiel de réutilisation des notions du cœur de l'ontologie. Une telle approche permet d'étendre assez facilement l'ontologie pour l'adapter aux besoins applicatifs à venir. Par exemple, nous avons ajouté, pour notre expérimentation, des relations basiques visant les éléments de contenu des documents, relations qui dénotent des jugements interprétatifs concernant la manière dont le contenu thématique est présenté par ces documents : *clarifie*, *exemplifie*, *démontre*... Finalement, nous proposons des relations conceptuelles liant les objets AV à des thèmes externes, ce qui permet la description du contenu proprement dit.

Considérant notre corpus de vidéos, nous avons examiné certaines des terminologies médicales existantes. L'une d'entre elles, l'ontologie MENELAS, décrit le domaine des pathologies coronariennes (Zweigenbaum & Consortium MENELAS, 1994), et contient beaucoup de concepts liés à la chirurgie cardiaque, le thème de notre corpus. Les concepts médicaux généraux de cette ontologie se retrouvent dans les connaissances formalisées dans d'autres ontologies du domaine, comme GALEN[2] (*General Architecture for Language and Nomenclatures*), un système dédié au développement d'ontologies dans tous les domaines médicaux, y compris les procédés chirurgicaux (Rector & Nowlan, 1993). Plutôt que d'aligner ces deux ontologies, nous avons introduit quelques équivalences de classes quand cela était nécessaire. D'une certaine façon, l'articulation entre l'ontologie AV et l'ontologie thématique est gérée par ces équivalences (`av:person` et `menelas :human_being`, par exemple). Dans la section suivante, nous allons montrer comment ces ressources sont utilisées pour décrire notre corpus.

# 3 Indexer les vidéos

## 3.1 Procédé d'annotation

Décrire un document AV implique de considérer des aspects documentaire (identifier les éléments qui constituent la structure logique du document) aussi bien que thématique (affirmer que ces éléments sont *à propos* de quelque chose). Distinguer l'ontologie AV des autres ontologies thématiques nous permet de considérer ces deux aspects.

Les concepts et les relations de l'ontologie de l'audiovisuel sont introduits dans les descriptions pour spécifier les liens entre les éléments de contenu documentaire au niveau de la connaissance. Ensuite, cette description conceptuelle peut être facilement liée à des méta-données strictement documentaires, exprimées avec un langage comme MPEG-7, en utilisant par exemple l'architecture décrite dans (Troncy, 2003). Les éléments documentaires sont ainsi décrits comme des ressources, classifiés sous des concepts AV.

[2] `http://www.opengalen.org`, 2001.

L'outil **SegmenTool**[3], illustré en figure 1, nous a permis de segmenter les documents AV et de créer les méta-données spécifiques à la description documentaire audiovisuelle.

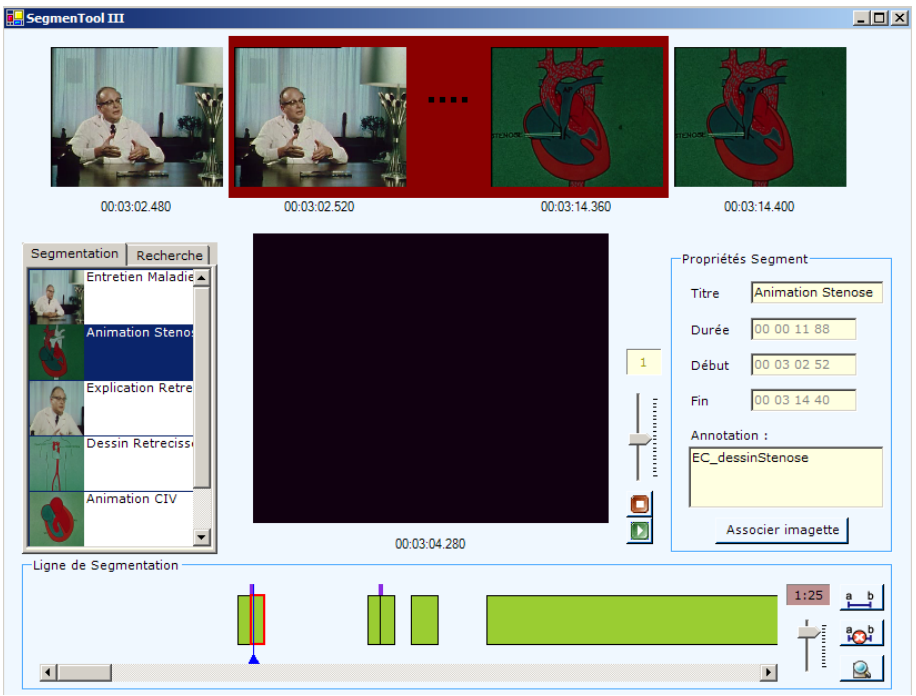



FIG. 1 – Utilisation de l'outil SegmentTool pour produire une structure documentaire

## 3.2 Annotation Conceptuelle

Pour faciliter le processus de description et rendre ses résultats plus cohérents, nous utilisons un *patron d'indexation* relationnel constituant un guide sur la manière dont concepts et relations sont utilisés Isaac *et al.* (2005). Ce patron est une construction relationnelle qui présente les notions ontologiques caractéristiques du domaine applicatif, employés dans leur contexte d'indexation typique. Cette structure est ensuite adaptée et spécialisée par l'indexeur pour rendre compte de son interprétation d'un élément documentaire donné. Cette vison est semblable à certaines approches d'indexation par formulaires ou *grilles*, à ceci près que nous sommes dans un cadre ontologique formalisé, où les concepts et relations présentés sont utilisables dans des raisonnements élaborés (voir la section 4). Et

[3]Cet outil a été développé par l'équipe DCA de la direction de la recherche de l'INA et a été partiellement financé dans le cadre du projet PRIAMM CHAPERON.

que, du fait même de l'utilisation de ces raisonnements, on peut se permettre une souplesse de description que n'autorisent pas les autres approches, condamnées à demeurer rigides pour garantir la cohérence de l'indexation et l'efficacité de son exploitation.

Pour trouver un patron d'indexation, il est nécessaire de se tourner vers les pratiques en cours dans le domaine d'application. Dans le cadre de notre expérimentation, nous nous sommes appuyés sur l'un des points de vue applicatifs du projet Opales, ainsi que sur les besoins d'analyse documentaire de l'INA dont une formalisation avait été proposée dans (Isaac & Troncy, 2004). Comme évoqué dans la figure 2, il faut décrire les éléments AV en leur assignant certaines valeurs de propriétés (par exemple, la manière dont ils sont produits) et en les décomposant d'un point de vue documentaire. Leur contenu doit également être indexé par l'assertion de relations avec des concepts du domaine, relations de nature strictement représentationnelle – ce qui est montré dans les vidéos – ou plus interprétatives – quelle est l'utilisation de ces représentations.

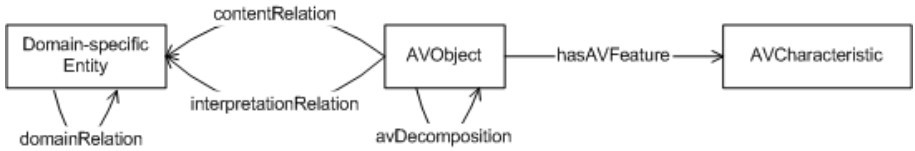



Fig. 2 – Le patron de description relationnel

Cette structure simple peut engendrer des descriptions extrêmement riches, puisqu'elle apporte une forme de récursivité, par l'intermédiaire des relations de décomposition relevant de l'audiovisuel ou du thème retenu. L'exemple de la figure 3 donne une idée des descriptions que nous cherchons à obtenir. Nous y avons mis en valeur la distinction entre les deux types de connaissances qui nous intéressent.

# 4 Rechercher et raisonner dans la base de connaissances

L'objectif de notre expérimentation est aussi de démontrer l'intérêt de l'utilisation de l'inférence sémantique dans des scénarios de recherche de contenu. Les assertions explicites que l'on trouve dans les descriptions telle que celle que nous avons montrée peuvent être complétées par des assertions *dérivées*, comme illustré en figure 4. Par exemple, si une séquence contient une séquence expliquant un sujet donné, on peut en déduire qu'elle aussi contribue à l'explication de ce sujet. On peut ainsi retrouver des objets AV qui font référence à des thèmes variés, même si lesdits thèmes n'apparaissaient explicitement que dans les éléments contenus dans ces objets. Ici, le système jugera notre documentaire

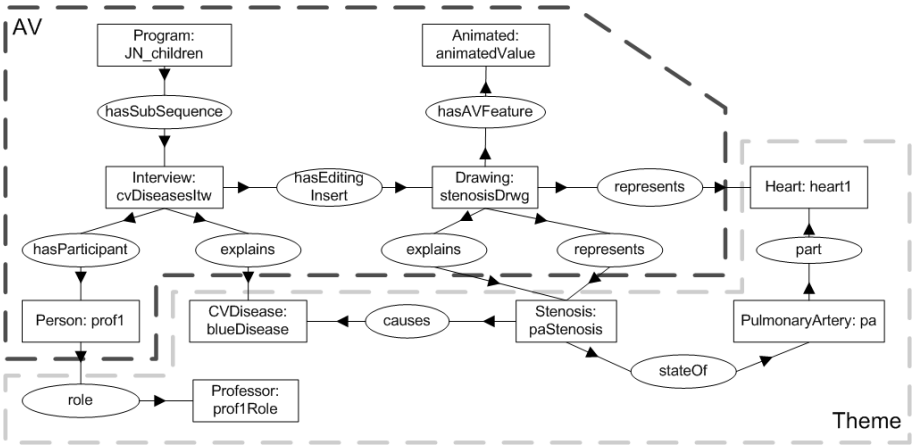



FIG. 3 – Un exemple de description

pertinent pour une requête telle que "trouver un programme qui explique une maladie et offre une représentation visuelle d'une de ses causes".

Pour bénéficier des pleines possibilités des techniques du web sémantique, nous devons étudier les manières d'encoder les connaissances de raisonnement, et de les articuler avec la gestion de la base de connaissances. Quels raisonnements pouvons nous rendre possibles, en fonction du langage de spécification d'ontologies retenu ?

Pour stocker et interroger les ontologies et les assertions, nous utilisons l'architecture **Sesame** (Broekstra *et al.*, 2002). Pour l'instant, cette architecture utilise RDF Schema comme langage de représentation d'ontologies, et offre les services de raisonnements conformes aux spécifications de la théorie des modèles RDF. Ceci autorise des inférences basiques, comme l'utilisation des liens de subsomption pour les concepts et les relations. Dans notre exemple, il est ainsi possible de trouver l'interview décrite dans les résultats d'une recherche de "séquences expliquant la maladie bleue", puisque dans l'ontologie de l'AV *Interview* spécialise *Séquence*.

Cependant, cela n'est pas suffisant pour l'exploitation que nous désirons, précise, et utilisant à la fois les propriétés des concepts et des relations présents dans les index. L'opportunité d'utiliser les possibilités des langages OWL – ou au moins celles du sous-ensemble décidable OWL-DL – semble en particulier séduisante. Avec OWL, on peut préciser qu'une *ExpertInterview* est exactement définie comme une interview où *au moins un* participant joue un rôle d'expert. Le concept *rôle d'expert* sera lui défini par le biais d'une équivalence de classes énumérant les rôles du domaine[4], qui peuvent être considérés comme dénotant une

[4]Dans notre cas, nous avons sélectionné parmi les spécialisations du concept *rôle* les concepts spécialisant *académique*, *professionnel* et *hospitalier*, à l'exception du rôle attribué à l'institution hospitalière elle-même, ce qui a été rendu possible en utilisant le constructeur OWL

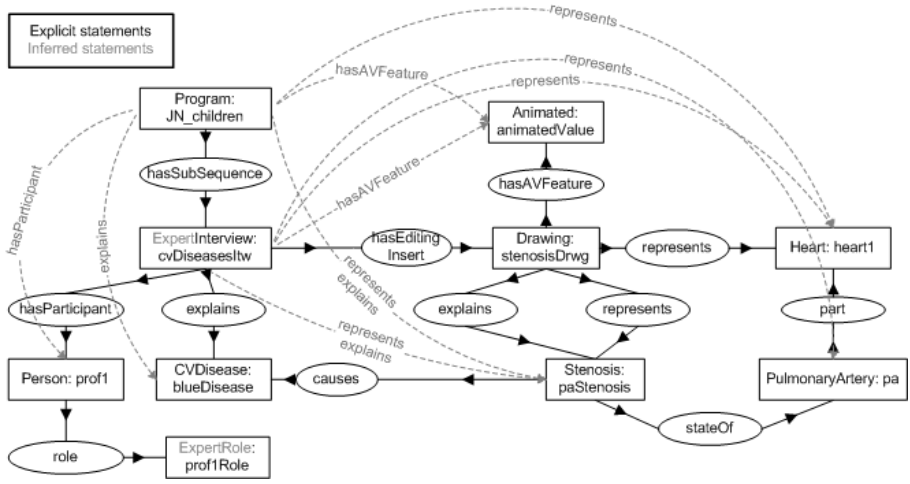



FIG. 4 – Index complété avec des connaissances inférées

certaine expertise pour l'application. Ainsi, on peut encoder les connaissances autorisant les inférences illustrées en grisé plein sur la figure 4, et répondre à des requêtes comme "émissions contenant des témoignages d'expert et expliquant une maladie cardio-vasculaire". Pour implémenter de tels raisonnements, on peut se tourner vers des raisonneurs OWL comme **BOR**[5] (Simov & Jordanov, 2002), qui a été intégré dans Sesame.

Et pourtant, cette solution ne satisfait pas complètement nos besoins. Comme nous ciblons des index riches en connaissances relationnelles, nous voudrions exploiter des connaissances de raisonnement exploitant ces relations. OWL-DL permet de spécifier des propriétés algébriques pour les relations. Ceci est clairement utile, mais on manque de possibilités pour encoder des connaissances plus générales, relatives en particulier à la composition des relations. Nous avons par exemple besoin de créer des règles comme `hasSubSequence(x,y)` $\wedge$ `represents(y,z)` $\Rightarrow$ `represents(x,z)`. On pourrait ainsi traiter une requête semblable à la précédente, mais demandant en sus que la séquence d'interview soit "illustrée par des images montrant l'objet concerné par la pathologie".

Ces préoccupations sont reconnues dans la communauté du web sémantique, et commencent à se voir traitées par des langages et des outils appropriés. Ainsi, SWRL (Horrocks *et al.*, 2004) constitue un pas vers le rapprochement entre les langages OWL et les règles logiques. Quelques-unes des possibilités considérées ici peuvent être implémentées *via* des cadres logiques décidables comme OWL-DLP (Grosof *et al.*, 2003) qui restreint OWL-DL tout en autorisant le recours

*complementOf*.

[5]BOR implémente en fait la sémantique du langage DAML+OIL, mais celle-ci est extrêmement proche de ce qui peut être spécifié pour les moteurs OWL.

à certains des éléments des programmes logiques. Perdre une partie de l'expressivité OWL – en particulier les restrictions existentielles dans les conditions nécessaires – peut se révéler gênant, mais la richesse des règles relationnelles possibles[6] nous fait préférer un tel choix.

Dans Sesame, une telle opportunité est implémentée dans un module d'inférence *sur mesure*, où les axiomes et les règles de RDFS sont complétées par ceux de OWL-DLP et des règles de raisonnement spécifiques aux ontologies exploitées. C'est une manière plutôt rigide de concevoir un tel système – les règles sont encodées au niveau de la spécification du raisonneur, et non dans l'ontologie elle-même – mais qui permet déjà de mettre en œuvre des raisonnements intéressants. On peut ainsi, de l'index de la figure 4, déduire les assertions grisées et pointillées qui viennent s'ajouter à celles issues des raisonnements OWL. On dispose alors de bien plus d'éléments dans la base de connaissances pour répondre aux requêtes. Le tableau 1 résume le nombre de triplets (explicites et inférés) contenus dans la base de connaissances Sesame pour notre expérimentation. La saturation de cette base est obtenue en utilisant les règles OWL-DLP complétées par une vingtaine de règles spécifiques à l'application, en majorité des règles de composition. Il est évident que l'enjeu de l'exploitation du raisonnement dans un système d'information va au-delà d'un gain mesuré quantitativement, puisqu'il s'agit de simuler, grâce à l'implémentation de connaissances de raisonnement propres à un domaine d'application, une partie des raisonnements[7] qui sont effectués par les recherchistes dans les systèmes documentaires actuels. Nous considérons cependant que, sur une expérimentation assez réduite, nous avons là, en complément de l'index complété de la figure 4, un indice assez révélateur de l'intérêt des mécanismes de raisonnement standardisés par le web sémantique.

| | **Triplets explicites** | **Triplets inférés** | **Total** |
|---|---|---|---|
| **Modèle RDF** | | | 129 |
| **Ontologie AV** | 5231 | 10810 | 16041 |
| **Ontologie Menelas** | 10534 | 26637 | 37171 |
| **Instances** | 276 | 1507 | 1783 |
| **Total** | 16041 | 38954 | 54995 |

TAB. 1 – Nombre de triplets (explicites and inférés) dans la base de connaissances *Sesame*. Le modèle RDF désigne les triplets définissant le langage de représentation lui-même.

[6] Ces règles peuvent d'ailleurs remplacer partiellement les éléments perdus de OWL-DL, comme dans le cas des restrictions existentielles dans des définitions par condition suffisante.

[7] Ces "raisonnements" constituent essentiellement des reformulations de requêtes, pour élargir ou préciser leurs résultats.

# 5 Travaux existants

Une partie des hypothèses et du travail présenté ici a déjà été mis en œuvre dans le cadre du projet Opales qui fournit un cadre général pour décrire manuellement le contenu de documents audiovisuels éducatifs en utilisant le formalisme des graphes conceptuels, et pour effectuer des recherches parmi les descriptions produites en utilisant un moteur d'inférence approprié. Cependant, la possibilité de faire référence explicitement à plusieurs ontologies pour produire les descriptions est clairement manquante dans cette architecture. L'expérimentation proposée ici constitue donc une évolution vers l'utilisation de langages et d'outils liés au web sémantique, profitant ainsi des nombreux efforts de recherche menés dans ce domaine. Les annotations sont ainsi exprimées en RDF, les ontologies sont représentées en OWL(DL), et toutes les ressources peuvent être distribuées et ré-utilisées dans d'autres applications.

Il y a relativement peu d'autres travaux concernant l'annotation de documents multimédias. Parmi ceux-ci, on peut citer (Hollink *et al.*, 2003) qui montre qu'adapter différents thésaurus à une application d'annotation d'œuvres d'art peut sensiblement améliorer à la fois le processus d'annotation sémantique pour les documentalistes et le processus de recherche. Nous avons en commun avec ce travail la possibilité d'utiliser plusieurs ressources ontologiques mais proposons un cadre d'indexation plus flexible et reposant de façon plus importante sur le raisonnement.

Le prototype nommé **Vannotea** a été développé pour permettre l'indexation collaborative, l'annotation et la discussion de contenu de documents audiovisuels à travers des réseaux à haut débit (Schroeter *et al.*, 2003). Cependant, cet outil se concentre sur la description des éléments documentaires composant le document, en représentant cette structure dans le format MPEG-7. L'annotation du contenu proprement dit reste principalement du texte libre, rendant les inférences sur la base des descriptions relativement limitées par rapport à celles proposées dans cet article.

Finalement, le projet MIAKT[8] (*Medical Imaging and Advanced Knowledge Technologies*) a pour but d'appliquer des technologies de représentation de connaissance et d'analyse intelligente de données à la résolution collaborative de problèmes dans le domaine de la surveillance et du diagnostique du cancer du sein (Dasmahapatra *et al.*, 2004). Ce projet se concentre sur l'annotation d'images médicales statiques, ce qui restreint l'étendue des descriptions possibles. Ainsi, les auteurs proposent un cadre relativement rigide pour la description des objets audiovisuels, qui diffère assez largement de celui proposé dans cet article, plus flexible et combinant des schémas de description.

[8] `http://www.aktors.org/miakt/`

# 6 Conclusion

Nous avons présenté dans cet article une expérimentation consistant à décrire le contenu de documents audiovisuels en utilisant des langages et des outils proposés pour le web sémantique. Nous avons exploré la manière dont une ontologie de l'audiovisuel et des ontologies liées à un domaine particulier pouvaient être articulées pour produire des descriptions pertinentes pour une application donnée. Nous avons également montré comment les connaissances de raisonnement peuvent être encodées et utilisées pour rendre les systèmes plus efficaces. Il est important de remarquer que les observations effectuées ici sont basées sur des cas réels d'utilisation. Nous ne détaillons ici qu'un début d'expérimentation et d'évaluation, mais les résultats tant qualitatifs que quantitatifs obtenus montrent la faisabilité et l'intérêt, pour les performances globales d'un système d'information, de l'utilisation des technologies développées dans le cadre du web sémantique pour décrire des documents audiovisuels.

Il est néanmoins important de noter que, comme dans toute approche de représentation de connaissances, un compromis entre l'expressivité d'une part, et les possibilités computationnelles en terme d'inférences d'autre part, doit être trouvé. Il faudra pour cela, et dans chaque cas applicatif, déterminer quels sont les besoins du système en matière d'implémentation de connaissances de raisonnement, et faire le choix d'outils appropriés. Pour cela, les propositions du web sémantique constituent un cadre idéal, puisque les techniques proposées, comme nous l'avons vu dans notre expérimentation, suivent une approche "incrémentale" en termes d'expressivité, de l'utilisation de connaissances représentationnelles élémentaires – RDF – aux systèmes plus élaborés et coûteux en complexité – OWL augmenté de règles relationnelles.

# Références

Bachimont B., Isaac A. & Troncy R. (2002). Semantic Commitment for Designing Ontologies : A Proposal. In A. Gómez-Pérez & V. R. Benjamins, Eds., *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, volume LNAI 2473, p. 114–121, Sigüenza, Espagne.

Broekstra J., Kampman A. & van Harmelen F. (2002). Sesame : a Generic Architecture for Storing and Querying RDF and RDF Schema. In I. Horrocks & J. Hendler, Eds., *1st International Semantic Web Conference (ISWC'02)*, volume LNCS 2342, p. 54–68, Sardaigne, Italie.

Dasmahapatra S., Dupplaw D., Hu B., Lewis H., Lewis P. & Shadbolt N. (2004). Facilitating multi-disciplinary knowledge-based support for breast cancer screening. *International Journal of Healthcare Technology and Management*.

Grosof B. N., Horrocks I., Volz R. & Decker S. (2003). Description Logic Programs : Combining Logic Programs with Description Logic. In *12th*

*International World Wide Web Conference (WWW'03)*, p. 48–57, Budapest, Hongrie.

Hollink L., Schreiber G., Wielemaker J. & Wielinga B. (2003). Semantic Annotation of Image Collections. In *Workshop on Knowledge Markup and Semantic Annotation*, Sanibel Island, Floride, USA.

Horrocks I., Patel-Schneider P. F., Boley H., Tabet S., Grosof B. N. & Dean M. (2004). SWRL : A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission (21 Mai). `http://www.w3.org/Submission/SWRL/`.

Isaac A., Bachimont B. & Laublet P. (2005). Indexation de documents AV : patrons de conception et d'utilisation. In *16$^{es}$ Journées Francophones d'Ingénierie des Connaissances (IC'2005)*, Nice, France.

Isaac A., Courounet P., Genest D., Malaisé V., Nanard J. & Nanard M. (2004). Un système d'annotation multiforme et communautaire de documents AV : Opales. In *Journée sur les Modèles Documentaires de l'Audiovisuel organisée dans le cadre de la Semaine du Document Numérique (SDN 2004)*, La Rochelle, France. `http://archivesic.ccsd.cnrs.fr/`.

Isaac A. & Troncy R. (2004). Designing and Using an Audio-Visual Description Core Ontology. In *Workshop on Core Ontologies in Ontology Engineering, 14$^{th}$ International Conference on Knowledge Engineering and Knowledge Management (EKAW'04)*, Whittlebury Hall, Northamptonshire, UK.

OWL (2004). Web Ontology Language Reference Version 1.0. W3C Recommendation (10 Février). `http://www.w3.org/TR/owl-ref/`.

RDF (2004). Ressource Description Framework Primer. W3C Recommendation (10 Février). `http://www.w3.org/TR/rdf-primer/`.

Rector A. L. & Nowlan W. (1993). The Galen Project. *Computer Methods and Programs in Biomedicine*, **45**, 75–78.

Schroeter R., Hunter J. & Kosovic D. (2003). Vannotea - A Collaborative Video Indexing, Annotation and Discussion System For Broadband Networks. In *Workshop on Knowledge Markup and Semantic Annotation*, Sanibel Island, Floride, USA.

Simov K. & Jordanov S. (2002). BOR : a Pragmatic DAML+OIL Reasoner. Deliverable 40, On-To-Knowledge Project.

Troncy R. (2003). Integrating Structure and Semantics into Audio-visual Documents. In D. Fensel, K. Sycara & J. Mylopoulos, Eds., *2$^{nd}$ International Semantic Web Conference (ISWC'03)*, volume LNCS 2870, p. 566–581, Sanibel Island, Floride, USA.

Troncy R. (2004). *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels*. PhD thesis, Université Joseph Fourier, Grenoble, France.

Zweigenbaum P. & Consortium Menelas (1994). MENELAS : An access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, **45**, 117–120.

# Adapter temporellement un document SMIL

Sébastien Laborie, Jérôme Euzenat, Nabil Layaïda

INRIA Rhône-Alpes, 655 avenue de l'Europe
38330 Montbonnot Saint-Martin, France
{Sebastien.Laborie,Jerome.Euzenat,Nabil.Layaida}@inria.fr

**Résumé** : Les récentes avancées technologiques permettent aux documents multimédia d'être présentés sur de nombreuses plates-formes (ordinateurs de bureau, PDA, téléphones portables…). Cette diversification des supports a entraîné un besoin d'adaptation des documents à leur contexte d'exécution. Dans [4], une approche sémantique d'adaptation de documents multimédia a été proposée et temporellement définie à l'aide de l'algèbre d'intervalles d'Allen. Cet article étend ces précédents travaux en les appliquant au langage de spécification de documents multimédia SMIL. Pour cela, des fonctions de traduction de SMIL vers l'algèbre de Allen (et inversement) ont été définies. Celles-ci préservent la proximité entre le document adapté et le document initial. Enfin, ces fonctions ont été articulées avec [4].

**Mots-clés** : Adaptation sémantique, Documents multimédia SMIL.

## 1 Introduction

Un document multimédia doit pouvoir être exécuté sur des plates-formes aux possibilités variées : téléphones, PDA, ordinateurs de bureau, lecteurs de salon… Afin de pouvoir tenir compte des possibilités de ces plates-formes, les documents sont transformés de telle sorte qu'ils puissent être rendus correctement sur la plate-forme cible. Dans un premier temps, il est nécessaire d'indiquer les caractéristiques temporelles des documents multimédia (§ 1.1). Puis, le problème de l'adaptation sera exposé (§ 1.2). Enfin, une approche sémantique d'adaptation sera présentée (§ 1.3).

### 1.1 Spécification temporelle d'un document multimédia

Cet article se consacre principalement à l'adaptation de documents multimédia selon leur dimension temporelle. Dans un document multimédia temporel, la présentation des objets multimédia est organisée dans le temps. Un tel document est présenté dans la Figure 1. Le temps est représenté sur l'axe horizontal. L'exemple proposé est une présentation d'une équipe de recherche contenant différentes régions composées d'objets graphiques qui peuvent être présentés simultanément. La première région affiche une image (Titre) et deux vidéos (Auteur et Demo). De plus, un Discours est joué durant une partie de la présentation. Chacun de ces objets est représenté par un segment dont

les extrémités de début et de fin correspondent respectivement au début et à la fin de leur présentation.

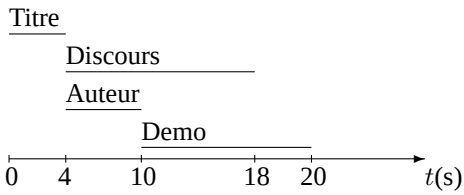



FIG. 1 – Dimension temporelle d'un document multimédia.

Le Titre débute à la seconde 0 et fini à la seconde 4. Le Discours débute à la seconde 4 et fini à la seconde 18, tandis que l'objet Auteur débute à la seconde 4 et fini à la seconde 10, et l'objet Demo débute à la seconde 10 et fini à la seconde 20. Une telle description est précise et quantitative car elle définit exactement les instants de début et de fin de chaque objet multimédia. Cette information est suffisante pour exécuter le document : à une représentation quantitative précise correspond une seule exécution possible du document (comprenant une référence temporelle fixe). Les documents multimédia ne sont pas toujours spécifiés précisément car il est plus commode pour l'auteur de laisser l'interprétation de la spécification à la machine tant que sa demande est clairement exprimée. Les spécifications non précises peuvent être exprimées par des relations qualitatives entre les objets multimédia. Par exemple, l'objet Discours *commence en même temps* que l'objet Auteur...

Il existe divers langages de spécification de documents multimédia avec différentes possibilités d'expression de la dimension temporelle : Magic [3] et Madeus [5] utilisent une restriction de l'algèbre d'intervalles d'Allen ; SMIL [1] exprime l'organisation des objets multimédia par des opérateurs parallèles ou séquentiels sur les intervalles. Cet article se consacrera à des documents exprimés avec le langage SMIL.

SMIL est un langage basé sur XML. Un document SMIL peut donc être vu comme un arbre dans lequel les nœuds, ou éléments, sont étiquetés. Deux catégories d'éléments sont identifiées : les objets multimédia (ex. texte, audio, vidéo, image) et les conteneurs de temps. Il existe deux principaux conteneurs de temps : un conteneur *séquentiel* (`seq`) et un conteneur *parallèle* (`par`). Chaque fils d'un conteneur de temps peut être un objet multimédia ou bien un autre conteneur de temps. Les objets multimédia sont les feuilles de l'arbre. De plus, un ensemble d'attributs comme `begin`, `end` ou `dur` peuvent être spécifiés sur chaque élément. Ceci permettant un meilleur contrôle de la synchronisation des éléments. Le document de la Figure 1 peut s'exprimer qualitativement en SMIL par le fragment de code de la Figure 2.

A partir d'une telle spécification, le système de présentation multimédia (ou le Player) calcule un plan (appelé scénario) qui peut être exécuté. Cette étape est appelée le formatage temporel.

```
<seq>
     <img src="Titre.jpg" dur="4s"/>
     <par>
          <audio src="Discours.au" dur="14s"/>
          <seq>
               <video src="Auteur.avi" dur="6s"/>
               <video src="Demo.avi" dur="10s"/>
          </seq>
     </par>
</seq>
```

FIG. 2 – Spécification d'un document SMIL

## 1.2 Adaptation de documents multimédia

Différents contextes de présentation multimédia introduisent différentes contraintes sur la présentation elle-même. Par exemple, les limitations de la bande passante entre le client et le serveur peuvent conduire le client à ne pas jouer deux vidéos au même instant. Les limitations dues à l'affichage peuvent mener à des contraintes similaires. D'autres types de contraintes peuvent être introduites par les préférences de l'utilisateur, la protection du contenu ou les capacités du terminal. Les contraintes imposées par le client sont appelées le profil.
Les profils peuvent être exprimés en terme de restriction sur le langage utilisé pour spécifier les documents cibles ou bien en termes de contraintes supplémentaires imposées sur les objets. Par exemple, si la plate-forme possède un écran avec des capacités limitées, il ne sera alors pas possible de présenter deux images simultanément sur le même écran.
Pour satisfaire ces contraintes, les documents multimédia doivent être adaptés avant d'être présentés. Plusieurs types d'adaptation peuvent être envisageables comme l'adaptation locale c'est-à-dire liée aux différents objets multimédia et l'adaptation globale c'est-à-dire liée à l'organisation du contenu de la présentation. Cet article se consacrera à ce deuxième type d'adaptation.
A partir du profil et du document initial, l'étape d'adaptation doit produire un document satisfaisant les contraintes exprimées dans le profil. Cette adaptation est généralement réalisée par un programme de transformation du document [9, 6]. Celle-ci peut être considérée comme implicite si l'on dispose de solutions alternatives. Par exemple, SMIL nous offre la possibilité d'utiliser une balise `switch`. Néanmoins, il est nécessaire de connaître à l'avance les différentes solutions d'adaptation. L'étape d'adaptation peut aussi être explicite c'est-à-dire en utilisant la sémantique du document. Les spécifications qualitatives sont centrales à ce deuxième type car elles permettent une adaptation efficace en fournissant plus de flexibilité. Dans ce qui suit, une approche d'adaptation sémantique de documents multimédia sera présentée.

## 1.3 Adaptation sémantique de documents multimédia

Les travaux exposés dans [4] précisent ce que doit être l'adaptation d'un document multimédia en utilisant une sémantique en théorie des modèles. Il consiste à interpréter un document comme l'ensemble de ses exécutions potentielles. Ainsi, une contrainte

liée à une plate-forme va restreindre l'ensemble de ces exécutions en ne retenant que les exécutions compatibles. Adapter c'est trouver ce sous-ensemble des exécutions ou, lorsqu'il est vide, trouver une exécution compatible suffisamment proche des exécutions initiales. Pour réaliser cela, l'ensemble des interprétations possibles est représenté à l'aide d'un graphe de contraintes résolues.

En ce qui concerne la dimension temporelle, l'algèbre de relations introduite par Allen [2] permet de représenter temporellement les documents. Celle-ci est composée de 13 relations : before ($b$), meets ($m$), overlaps ($o$), starts ($s$), during ($d$), finishes ($f$), equals ($e$), finished-by ($fi$), contains ($di$), started-by ($si$), overlapped-by ($oi$), met-by ($mi$) et after ($bi$). La spécification SMIL de la Figure 2 peut se représenter par le graphe de relations d'Allen de la Figure 3 (gauche).

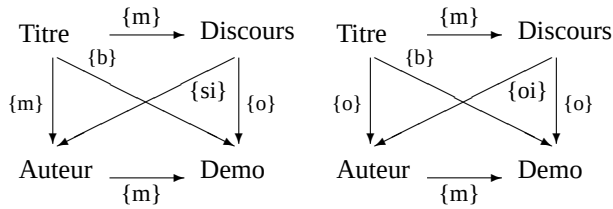



FIG. 3 – Graphe de relations de la spécification de la Figure 2 (gauche) et graphe de relations adapté (droite).

Lorsque le graphe de relations ne respecte pas les contraintes d'adaptation, il est alors nécessaire de l'adapter. Par exemple, supposons que l'on dispose du profil suivant : ne pas exécuter plus d'une vidéo en même temps et ne pas faire débuter plus d'un objet transmissible en flux ("streamable") au même instant. Il est clair que le graphe de relations de la Figure 3 (gauche) ne respecte pas les contraintes d'adaptation car l'objet Discours commence en même temps que l'objet Auteur à cause de la relation started-by ($si$). Notre approche sémantique d'adaptation permet de trouver un graphe de relations adapté (c.a.d respectant les contraintes d'adaptation) proche du graphe initial. Pour cela, des relations de proximité entre relations temporelles sont introduites et représentées dans un graphe de voisinage (Figure 4). Il est alors possible de calculer une distance entre relations ainsi qu'une distance entre graphes de relations. Adapter c'est trouver une distance minimale entre le graphe de relations initial et le graphe de relations adapté respectant les contraintes d'adaptation de la plate-forme cible. La Figure 3 (droite) présente ce graphe de relations adapté. Celui-ci est à une distance de 2 du graphe de relations initial.

Un document SMIL exprimé qualitativement doit pouvoir être adapté selon le même principe. Pour cela, il faut exprimer le réseau de contraintes entre les différents objets multimédia de la présentation et l'adapter en fonction du profil. Il est alors nécessaire de développer l'application du graphe de relations proposé vers ce langage (§2). Cela nécessite de projeter chacune des composantes de SMIL dans ce graphe et de transférer le résultat du graphe de relations vers SMIL (§3). Mais ceci soulève des problèmes de nature théorique. En effet, la transformation du graphe de contraintes vers la représentation SMIL ne garantit pas que le document adapté sera le plus proche possible du document initial. Pour cela, il faut étendre la représentation du document et préserver

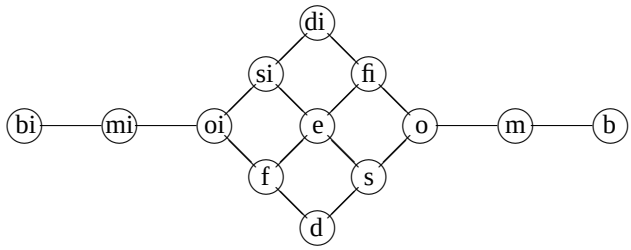



FIG. 4 – Graphe de voisinage de relations d'Allen.

suffisamment d'information concernant le document initial. Mais il est alors indispensable de réinsérer le résultat de l'adaptation dans cette représentation afin que le résultat reste cohérent et minimal (§4).

# 2 Principes d'adaptation de documents SMIL

L'étape d'adaptation présentée précédemment, si celle-ci est efficace, ne s'applique pas à un langage de spécification de documents multimédia particulier. Celle-ci doit être précisée pour chaque langage. Nous présentons son adaptation à SMIL [1].
La manière la plus naturelle d'utiliser l'approche d'adaptation précédente sur des documents ($S$) édités en SMIL, consisterait à prendre l'équivalent en algèbre d'Allen ($\alpha$), l'adapter et le traduire en SMIL ($\beta$, voir Figure 5).

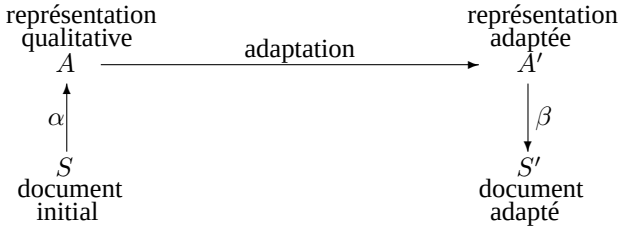



FIG. 5 – Stratégie générale.

Traduire un document SMIL en algèbre d'Allen n'est, en fait, pas une tâche très difficile. Cependant, cette traduction est en générale non injective et il est donc impossible de retraduire le document en retour car le résultat risque d'être trop éloigné du document initial.
Par exemple, soient $S_1$ et $S_2$ deux spécifications SMIL telles que :
$S_1$ = `<SEQ> <A/> <B/> </SEQ>` et $S_2$ = `<PAR> <A id="A"/> <B begin="A.end"/> </PAR>`, alors $\alpha(S_1) = \alpha(S_2) = A\{meets\}B$. Ces deux spécifications comportent des indications spécifiées par l'auteur que n'indique pas le graphe de relations. Dans notre cas, il s'agit des types des balises et des attributs temporels. Il est donc nécessaire de conserver ces informations. Ceci peut être effectué en définissant un arbre pour conserver la structure ainsi que le contenu du document initial. Cet arbre est construit par la fonc-

tion $\alpha$ et utilisé par la fonction $\beta$. De plus, il est nécessaire d'injecter les informations d'adaptation dans cette structure de façon consistante et minimale. Ceci est réalisé par la fonction $\gamma$. Les différentes étapes sont définies dans la Figure 6.

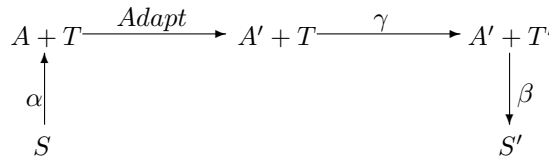



FIG. 6 – Ajustement structurel après l'adaptation.

Le but de l'adaptation est, comme toujours, de trouver la spécification qui satisfait les contraintes d'adaptation. Pour cela, il ne faut pas définir une fonction de SMIL vers Allen, mais plutôt une fonction minimisant la distance $d$ entre la spécification initiale et le document adapté.

**Définition 1 (Minimalité)**
*L'adaptation d'un document SMIL est minimale ssi* $\forall S \in \mathcal{S}$, $\beta\ o\ \gamma\ o\ Adapt\ o\ \alpha(S)$ *est à une distance minimale de* $S$ *à partir de tous les documents SMIL satisfaisant les contraintes d'adaptation.*

Dans le cas d'une spécification qui ne nécessite pas d'adaptation, il est important que la traduction ne change rien à la spécification. Ceci entraînant la propriété importante de neutralité.

**Définition 2 (Neutralité)**
$\beta\ o\ \alpha = id$

Il est à noter que ces propriétés sont totalement indépendantes du langage initiale et qu'elles doivent être vérifiées pour n'importe quel autre langage que SMIL.
Dans la partie 3, les fonctions de traduction satisfaisant ces propriétés vont être introduites. La partie 4 définira la fonction $\gamma$.

# 3 Fonctions de traduction

Une première solution consiste à encoder la structure du document SMIL en sortie de la fonction $\alpha$. Cette structure correspond exactement à la structure d'arbre du document comme il est possible de l'extraire à partir de n'importe quel document XML (voir Figure 7, gauche). Les objets multimédia se trouvent sur les feuilles de l'arbre.
Il n'est toujours pas possible de satisfaire $\beta\ o\ \alpha = id$ car l'arbre SMIL ne contient aucune information sur les attributs temporels d'un objet multimédia particulier. Par conséquent, la fonction $\beta$ n'est pas en mesure de retourner ce type d'information.
En effet, soient $S$ et $S'$ deux spécifications avec :
$S =$ `<par> <A id="A"/> <B begin="A.end"/> </par>`
et $S' =$ `<par> <A end="B.begin"/> <B id="B"/> </par>`,

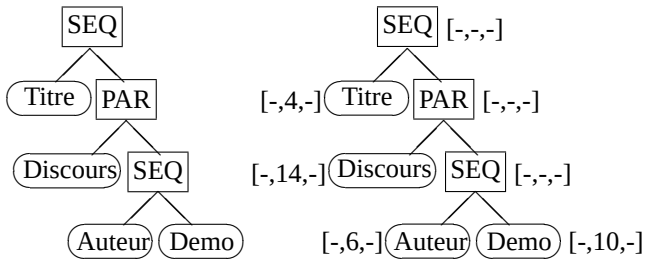



FIG. 7 – Arbre SMIL (gauche) et structure SMIL (droite) correspondant à la spécification de la Figure 2.

si $\alpha(S) = \alpha(S')$, alors $\beta(\alpha(S)) = \beta(\alpha(S'))$ et par conséquent le résultat de l'application de la fonction $\beta$ ne reflétera pas la structure initiale.

L'arbre SMIL doit donc être étendu avec des indications sur les attributs temporels utilisés dans le document initial. Pour cela, il est nécessaire d'introduire une structure SMIL qui associe à chaque nœud un index temporel. Ces index sont des triplets $[b, d, e]$ tel que $b$, $d$ et $e$ correspondent respectivement aux valeurs des attributs temporels `begin`, `dur` and `end`. Si l'attribut n'a pas de valeur, cela est noté par le symbole "-".

**Définition 3 (Structure SMIL)**
*Une structure SMIL $T = \langle E_N, E_O, r, R, \lambda, S\rangle$ avec $E_N$ un ensemble de nœud étiqueté par PAR ou SEQ, $E_O$ un ensemble de nœud correspondant aux objets multimédia, $r \in E_N$ un nœud racine, $R \subseteq E_N \times (E_N \cup E_O)$ un ensemble d'arc tel que $\langle E_N \cup E_O, R, r\rangle$ forme un arbre, $\lambda$ une fonction d'étiquetage $(E_N \cup E_O) \rightarrow (N \cup \{-\})^3$ qui associe à chaque nœud $n$ un index temporel $\lambda(n)$, et $S \subset (E_N \cup E_O)^2$ un ordre total sur les nœuds.*

La Figure 7 (droite) présente la structure SMIL correspondant à la spécification SMIL de la Figure 2.

De plus, la fonction $\alpha$ doit extraire la structure relationnelle des objets multimédia et l'encoder dans un graphe de relations sur lequel sera appliquée l'opération d'adaptation définie précédemment. Il est alors nécessaire d'extraire tous les objets multimédia et d'identifier les relations d'Allen entre chaque couple d'objets. Les informations portées par les conteneurs de temps et les attributs temporels nous permettent de définir ces relations. Le graphe complet de relations extrait est celui de la Figure 3 (gauche).

La fonction $\beta$ peut maintenant extraire l'arbre SMIL à partir de la structure SMIL et assigner les attributs temporels grâce aux valeurs des index.

Le résultat de neutralité est atteint par le couple de fonction définie.

Avec ces deux fonctions, il est maintenant possible de considérer la stratégie d'adaptation de document SMIL de la Figure 6. La fonction $\gamma$ va être définie ci-après.

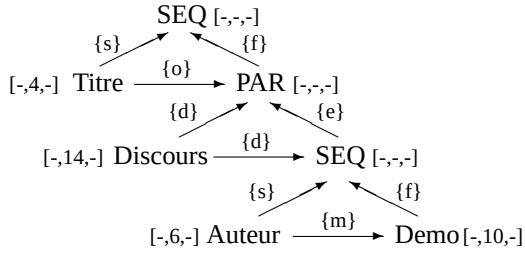
# 4 Maintenir la cohérence entre relations et structure

Comme présenté dans la Figure 6, à partir du document SMIL $S$, $\alpha(S) = \langle A, T \rangle$. Il est maintenant possible d'appliquer l'adaptation sur $A$ fournissant $\langle Adapt(A), T \rangle$. La suite logique correspondrait à appliquer la fonction $\beta$ à la paire résultante. Cependant, $\beta$ nécessite une structure SMIL cohérente avec le graphe de relations. L'étape d'adaptation introduit néanmoins des incohérences et en conséquence $\beta$ pourrait produire des documents SMIL illégaux. Par exemple, la structure SMIL de la Figure 7 (droite) est incohérente avec le graphe de relations adapté de la Figure 3 (droite), à cause de la relation *overlaps* ($o$) entre Titre et Auteur, et le conteneur de temps racine `seq` qui contraint les objets à être joués en séquence. Il est alors essentiel de transférer les informations d'adaptation dans la structure SMIL.

## 4.1 Restaurer la cohérence

Comme il est possible de constater avec l'exemple ci-dessus, l'incohérence peut venir des conteneurs de temps non adaptés aux relations d'Allen du graphe de relations mais elle peut aussi venir des index temporels. La fonction $\gamma$ sera alors composée de deux étapes : restaurer la cohérence des conteneurs de temps (`CohérenceBalise`) et restaurer la cohérence des index temporels (`CohérenceIndex`).
Pour restaurer la cohérence, il est aussi nécessaire de raisonner conjointement avec le graphe de relation $A$ et la structure SMIL $T$. La notion d'intervalles de référence [2] permet d'introduire ces relations dans la structure SMIL. Au lieu d'introduire le graphe de relations complètement, seules les relations entre nœuds et leur ancêtre et nœuds et leurs frères sont ajoutées à $T$.

**Définition 4 (Structure SMIL étendue)**
*Une structure SMIL étendue est une structure SMIL dans laquelle le graphe complet de relations entre frères est ajouté et tous les arcs sont étiquetés par les contraintes du graphe de relations.*




`CohérenceBalise`($n$) =
  $E := \{n'; \langle n, n' \rangle \in R\}$ *//enfant de n*
    For each $m \in E$
      `Coherence`($m$)
    If $type(n) =$ `seq` and $\forall m, p \in E$;
      $m \neq p \wedge$
      $relation(m, p) \not\subseteq \{b, m, mi, bi\}$
        MOD-N($n$)
    If $type(n) =$ `seq`
      Order($E$) *//Utiliser SWITCH*

FIG. 8 – Structure SMIL étendue (gauche) et procédure `CohérenceBalise` (droite).

A partir de la structure SMIL de la Figure 7 (droite) et du graphe de relations de la Figure 3 (droite), on peut construire la structure SMIL étendue de la Figure 8 (gauche). L'incohérence peut facilement être détectée à l'intérieur de cette nouvelle structure. En

particulier, le type du conteneur de temps racine stipule que tous ses fils sont exécutés séquentiellement alors qu'il existe une relation *overlaps* ($o$) entre des nœuds fils.
La procédure `CohérenceBalise` qui restaure la cohérence des conteneurs de temps d'une structure SMIL étendue est définie dans la Figure 8 (droite). Celle-ci utilise des opérations d'édition, comme MOD-N et SWITCH, qui seront définies dans la partie suivante.
Ensuite, la procédure `CohérenceIndex` qui restaure la cohérence des index temporels d'une structure SMIL étendue est appliquée. Celle-ci calcule une solution quantitative pour chaque index temporel et insère ce résultat dans la structure SMIL étendue si cela est nécessaire.
Une fois ces procédures appliquées sur la structure SMIL étendue, il est trivial de retourner une nouvelle structure SMIL. La Figure 9 (gauche) montre la structure SMIL adaptée de la structure SMIL de la Figure 7 (droite).

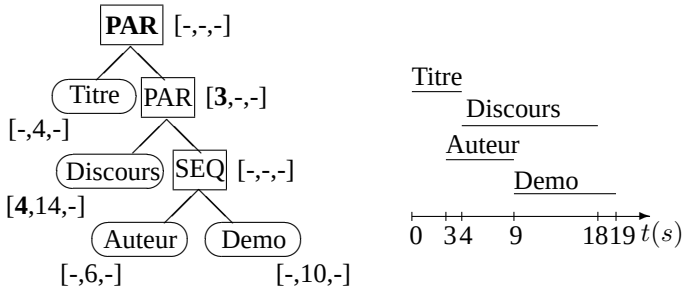



FIG. 9 – Structure SMIL adaptée (gauche) et timeline correspondante (droite).

En résumé, la fonction $\gamma$ construit une structure SMIL étendue, applique la procédure `CohérenceBalise` puis la procédure `CohérenceIndex` et restaure une structure SMIL adaptée. Le résultat est donc une structure SMIL cohérente sur laquelle $\beta$ peut être appliquée.

## 4.2 Minimalité

Dans le but de montrer que la fonction $\gamma$ produit un résultat minimal, une distance entre structure SMIL doit être définie. La notion de distance d'édition sera utilisée laquelle sera basée sur quelques opérations.

**Définition 5 (Opérations d'édition)**
*Cinq opérations d'édition sont possibles sur une structure SMIL :*

**ADD(**$n$**,**$m_1$**,…** $m_p$**)** *Ajoute un nœud $n'$ dans $E_N$ ainsi qu'un ensemble d'arcs $\langle n, n' \rangle$ et $\langle n', m_i \rangle$ à $R$ (les anciens arcs entre $n$ et $m_i$ sont supprimés).*

**DEL(**$n$**)** *Supprime le nœud $n$, tous ses arcs sont supprimés et tous ses fils sont attachés à son père (et introduit dans le même ordre).*

**MOD-M(**$n$**,** $m$**,** $[b, d, e]$**)** *Modifie l'index temporel du nœud $n$.*

**MOD-N(**$n$**)** *Change le type du conteneur de temps $n$ (de `seq` en `par` et vice versa).*

**SWITCH(**$n_i$**,** $n_{i+1}$**)** *Permute l'ordre des nœuds $n_i$, $n_{i+1}$ (fils du même conteneur).*

Il est à noter qu'il n'est pas possible d'ajouter ou bien de supprimer des objets multimédia du document SMIL car l'adaptation définie dans [4] n'utilise pas ces opérations. La fonction $\gamma$ définie précédemment n'utilise que les opérations d'édition MOD-M, MOD-N et SWITCH. De plus, on n'ajoute ni n'efface aucun nœud. Il est important d'indiquer que l'on peut utiliser la suppression pour simplifier la structure du document SMIL (par exemple effacer un `par` dans un autre `par` : ceci pourrait être appliqué à la Figure 9 gauche) mais cela mènerait vers des documents trop éloignés.
Il est clair à partir de ces définitions que n'importe quel couple de structures SMIL sur le même ensemble d'objets multimédia $O$ peut être construit à partir des opérations d'édition.

**Proposition 1 (Accessibilité)**
*A partir de structures SMIL $T$ et $T'$ quelconques basées sur le même ensemble d'objets multimédia $O$, il est possible de transformer $T$ en $T'$ en appliquant les opérations d'édition de la définition 5.*

**Définition 6 (Distance d'édition entre structures SMIL)**
*Soient deux structures SMIL $T$ et $T'$, leur distance d'édition $\delta(T, T')$ est la somme minimale des poids de chaque opération d'édition générant $T'$ à partir de $T$. Les poids de DEL et ADD sont de 4, celui de SWITCH est de 3, celui de MOD-N est de 2, et celui de MOD-M est de 1.*

Les poids ont été choisis pour préserver au maximum la structure du document spécifiée par l'auteur. Les poids les plus élevés sont associés aux modifications de la structure, puis aux modifications de type de balise, et enfin aux modifications d'index. La distance entre les structures SMIL de la Figure 7 (droite) et 9 (gauche) est de 4 (2$\times$ MOD-M + MOD-N).
Il est enfin possible d'établir le résultat attendu : $\gamma$ est la transformation minimale pour restaurer la cohérence car chaque opération d'édition effectuée est la moins coûteuse et est inévitable.

**Proposition 2 (Minimalité)**
$\gamma(\langle A', T\rangle) = \langle A', T'\rangle$ *avec* $\delta(T, T')$ *minimale.*

# 5 Résultats expérimentaux

Les travaux présentés précédemment ont été implémentés. La vue globale du système d'adaptation est présentée dans la Figure 10.
L'auteur peut éditer son document multimédia SMIL à l'aide de l'éditeur situé au centre de la figure. Le système comporte également une vue de la structure SMIL (en haut à droite), de la timeline du document (en bas à droite) ainsi que du graphe de relations comportant les relations sur l'algèbre d'intervalles d'Allen (à gauche). Dans un second temps, il est possible d'indiquer au système des contraintes à appliquer au document (c.a.d spécifier le profil). Si le document ne nécessite pas d'adaptation alors le document adapté est identique au document initial. Dans le cas contraire, s'il existe plusieurs solutions d'adaptation le système présente toutes les solutions minimales. Une

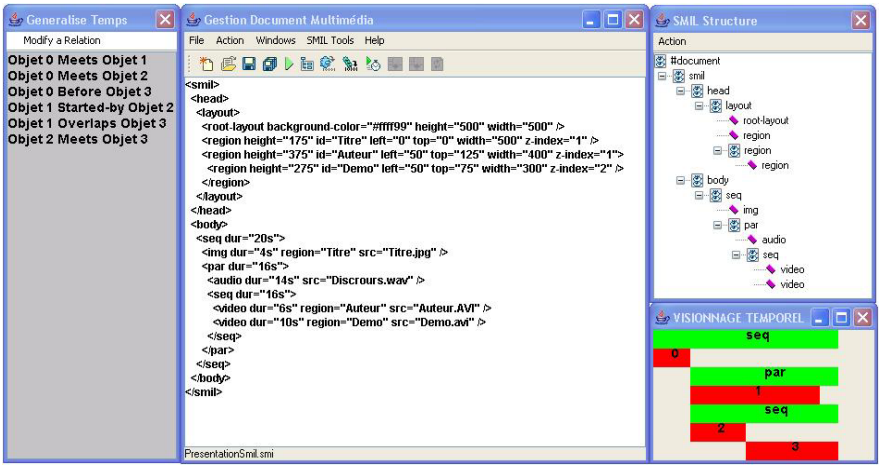



FIG. 10 – Vue globale du système d'adaptation.

fois la solution choisie, le système produit un document SMIL adapté et l'utilisateur peut ensuite exécuter le document multimédia pour le visualiser.

# 6 Limitations

Ces travaux sont limités à la dimension temporelle, alors que l'adaptation peut tirer parti des autres dimensions.
Une autre limitation est que $\gamma$ est minimale mais $\gamma \; o \; Adapt$ ne l'est pas. Il est alors nécessaire de contraindre la fonction $Adapt$ pour obtenir une entière minimalité. De telles contraintes doivent être utilisées lorsque plusieurs solutions d'adaptation sont possibles. Dans un tel cas, on doit décider quelle solution est la plus proche.

# 7 Travaux antérieurs

D'autres travaux comme [8, 7] portent sur l'adaptation de documents multimédia. Ceux-ci se concentrent sur l'utilisation d'une spécification particulière du document pour générer des documents SMIL adaptés plutôt que d'adapter des documents SMIL existants.

# 8 Conclusion

Cet article applique les précédents travaux d'adaptation sémantique de documents multimédia [4] à la dimension temporelle des documents SMIL. L'adaptation de documents multimédia SMIL proposée assure que l'adaptation est appliquée lorsque cela

est nécessaire et, si tel est le cas, est minimale. La propriété de minimalité est essentielle pour l'auteur souhaitant reconnaître la spécification de son document, mais elle est aussi nécessaire lorsque plusieurs étapes d'adaptation (selon différents critères) sont enchaînées. Pour cela, il a fallu définir des fonctions de traduction de SMIL vers l'algèbre d'intervalles d'Allen. Ces fonctions possèdent la propriété de neutralité lorsque l'adaptation n'est pas nécessaire. Nous avons défini des post-procédures à l'adaptation produisant un document le plus proche possible de la spécification initiale.
Enfin, une implémentation de nos travaux a été réalisée. L'utilisateur peut éditer son document SMIL, définir des contraintes à appliquer à son document, visualiser le document SMIL adapté et l'exécuter.

# Références

[1] (2001). *Synchronized Multimedia Integration Language (SMIL 2.0) Specification*. W3C. http ://www.w3.org/TR/smil20/.

[2] ALLEN J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, **26**(11), 832–843.

[3] DALAL M., FEINER S., MCKEOWN K., PAN S., ZHOU M. X., HOLLERER T., SHAW J., FENG Y. & FROMER J. (1996). Negotiation for automated generation of temporal multimedia presentations. In *ACM Multimedia conference*, p. 55–64, Boston (MA US).

[4] EUZENAT J., LAYAÏDA N. & DIAS V. (2003). A semantic framework for multimedia document adaptation. In *Proc. 18th International Joint Conference on Artificial Intelligence (IJCAI), Acapulco (MX)*, p. 31–36.

[5] JOURDAN M., LAYAÏDA N., ROISIN C., SABRY-ISMAÏL L. & TARDIF L. (1998). Madeus, an authoring environment for interactive multimedia documents. In *6th ACM Multimedia conference*, p. 267–272, Bristol (UK).

[6] LEMLOUMA T. & LAYAÏDA N. (2001). The negotiation of multimedia content services in heterogeneous environments. In *Proc. 8th International Conference on Multimedia Modeling (MMM01)*, p. 187–206, Amsterdam (NL).

[7] SCHERP A. & BOLL S. (2004). mobileMM4U – framework support for dynamic personalized multimedia content on mobile systems. In *Proc. Techniques and Applications for Mobile Commerce (TaMoCO), Essen (DE)*.

[8] VAN OSSENBRUGGEN J., CORNELISSEN F., GEURTS J., RUTLEDGE L. & HARDMAN L. (2000). *Cuypers : a semi-automatic hypermedia generation system*. Rapport interne INS-R0025, CWI, Amsterdam (NL).

[9] VILLARD L. (2001). Authoring transformations by direct manipulation for adaptable multimedia presentations. In *Proc. ACM Symposium on Document Engineering (DocEng'01)*, p. 125–134, Atlanta (US).