Structuration automatique de talk shows télévisés

Vallet Félicien

Thèse

présentée pour obtenir le grade de docteur
de l'École TÉLÉCOM ParisTech

Spécialité : Signal et Images

Félicien Vallet

Structuration automatique de talk shows télévisés

Soutenue le 21 septembre 2011 devant le jury composé de :

Guillaume Gravier	Rapporteurs
Philippe Joly	
Bernard Mérialdo	Président
Jenny Benois-Pineau	Examinateurs
Sid-Ahmed Berrani	
Gaël Richard	Directeur de Thèse
Slim Essid	Encadrants
Jean Carrive	

Encore une sale affaire pour le commissaire Bougret et l'inspecteur Charolles...

Remerciements

Il est agréable de débuter ce manuscrit en remerciant les personnes qui, de près ou de loin, m'ont permis de mener à bien mon travail de thèse. Quelle qu'ait été leur implication au cours de ces quatre années, je leur adresse à toutes de vifs remerciements, tout en espérant n'en avoir oublié aucune.

En premier lieu, je tiens à remercier mes trois encadrants de thèse pour m'avoir conduit à bon port à l'occasion de cette Odyssée. S'il n'a pas duré dix ans, le parcours fut, comme pour Ulysse, semé d'embûches. Cependant, à chaque fois que les vents semblaient m'être défavorables, j'ai pu me reposer sur cette formidable équipe aux talents complémentaires pour continuer mon chemin. Je remercie ainsi mon directeur de thèse Gaël Richard qui a fermement tenu le rôle de vigie pour me permettre de garder le cap quand mes yeux n'y voyaient plus très clair. Je remercie également Jean Carrive, mon encadrant à l'Ina, qui m'a épaulé, encouragé et apporté un soutien aussi chaleureux qu'indéfectible. Enfin, je remercie très vivement le bouillonnant Slim Essid, qui m'a secondé dans toutes mes aventures. Enthousiaste, enjoué et optimiste (certains diraient trop mais est-ce là un défaut ?), il a été un compagnon de route extrêmement précieux. Mon seul souhait désormais est que cette fine équipe ait à nouveau l'occasion d'être réunie pour permettre à un autre de traverser sereinement les eaux tumultueuses du doctorat.

Je tiens également à adresser des remerciements sincères à Messieurs Guillaume Gravier et Philippe Joly pour avoir accepté de rapporter mes travaux de thèse et par la même occasion à tous les membres du jury : Madame Jenny Benois-Pineau et Messieurs Sid-Ahmed Berrani et Bernard Merialdo.

Mon travail a été l'occasion d'effectuer de nombreuses collaborations et par là même des rencontres humaines. Je pense en particulier à Zaïd Harchaoui (Télécom ParisTech), Simon Bozonnet et Nicholas Evans (EURECOM), et Florian Kaiser (TU Berlin).

Qui plus est, si effectuer une thèse CIFRE offre l'avantage de cotoyer conjointement les univers académique et industriel, cela présente également l'inconvénient (relatif) de multiplier par deux le nombre de personnes à remercier. En ajoutant à cela un nombre important de changements de bureau, recrutements, arrivées de nouveaux doctorants — autant à Télécom ParisTech qu'à l'Ina — on appréciera l'abondance d'individus dont j'ai eu la chance de croiser le chemin...

Au sein de l'Ina je tiens ainsi à remercier le machiavélique binôme formé par Thomas et Jérôme, de même que Valentine, Mattéo, Quentin, Fabrice, Hervé, Ludivine, Jérémy, les deux Sébastien, Matthew, Damien, Benjamin, Pierre, Clément, Laurent, Jean-Pascal, Rachid, Victor, Claude, Marie-Luce, Olivier, Odile, Chantal, Véronique, Philippe et Daniel.

De même à Télécom ParisTech je remercie de leur présence les compères du bureau DA408, Benoît et François, jamais les derniers pour vider un verre, ainsi que Kévin, Kristoffer, Cyril, Romain, Mounira, Mathieu, Jean-Louis, Valentin, Nancy, Benoît, Manuel, Rémi, Thomas, Antoine, Jérémy, Cédric, Roland, Bertrand, Jacques, Maurice, Yves, Fabrice et Sophie-Charlotte.

Et puis, il y a les autres. Ceux avec qui l'on vit, que l'on voit le soir, le week-end ou pendant les vacances : les amis. Je tiens donc à remercier dans le plus parfait désordre la galaxie des musiciens : les deux Nicolas, Matthieu, Greg, Mathias, Sébastien, Marc et la nébuleuse qui gravite tout autour : Pierre, Viking, Jef, Caroline, Stéphanie, Évelyne, Gilles, Ida, Ashley, Sandra, Isabel, Franck, Vincent, Laure, Sara, Etienne, Camille, Charlotte, etc. Il y a aussi les vieux copains, loin des yeux mais pas du coeur : Alexis, Anthony, Ant', John, Gabriel, Clare, Samuel et Manami.

Enfin, j'ai une pensée particulière pour ma famille : mon papa, ma maman et mes soeurs Nina et Flore. Je vous remercie de votre présence et vous témoigne ici toute mon affection.

Mais si la liste s'arrêtait là, il manquerait une personne indispensable à ma vie. Aussi, merci Céline d'avoir été là, dans les moments de doute comme dans ceux de joie. Merci de m'avoir encouragé, sollicité, questionné. Merci pour ton rire, le meilleur remède aux pensées sombres que je connaisse. Et surtout, merci pour ton amour.

Table des matières

Table des figures

Liste des tableaux

Introduction

Dialogue

Dans les bureaux de leur rédaction, X et Y évoquent la constitution d'un documentaire sur les écrivains de polars et autres romans noirs...

« Ce qui serait bien, ce serait d'avoir des témoignages et extraits vidéo d'auteurs qui expliquent leurs œuvres, donnent des éclairages nouveaux, parlent de leurs sources d'inspiration. Tu vois ? Des interviews de Raymond Chandler, Simenon, Fred Vargas par exemple.
— Oui, et ça me fait d'ailleurs penser qu'on doit pouvoir mettre la main sur des explications de James Ellroy sur sa fascination pour les crimes crapuleux. Comme c'est l'objet d'un de ses livres [NDLR : *Ma part d'ombre*] , il y a obligatoirement des archives là-dessus ! Ça doit se trouver, non ?
— Moi je me rappelle avoir vu il y a quelques temps une discussion assez animée entre Manchette, ADG et Léo Malet. C'était un vieux programme des années 80, présenté par Bernard Pivot...
— *Apostrophes* ?
— Oui ça doit être ça. Le problème c'est que je ne me souviens pas du tout de l'année de diffusion et cette émission a été à l'antenne au moins quinze ans. Sans parler de la question des droits d'exploitation, récupérer les trois minutes qui nous intéressent, ça va être coton !
— Hum... Comme dirait l'autre, les indices sont plutôt maigres patron ![1] »

Contexte général

Abandonnons ici nos deux journalistes pour réfléchir aux questions soulevées par cette discussion fictive. En effet, celle-ci met en évidence les problématiques modernes de conservation du patrimoine numérique, d'augmentation des volumes de données, de navigation dans les contenus multimedia, d'indexation des documents, de droits d'exploitation, etc. C'est d'ailleurs avec ce type de constat que débutent généralement les premiers paragraphes des articles relatifs aux domaines de l'image, de l'audio ou de la vidéo. Ceci n'est pas étonnant tant les difficultés rencontrées s'avèrent importantes, donnant ainsi naissance à de multiples orientations de recherche.

L'indexation automatique est une des grandes thématiques des dix dernières années. Elle peut se définir comme l'utilisation de méthodes logicielles permettant d'établir un index pour un ensemble de documents et faciliter ainsi l'accès ultérieur à leur contenu. Ce domaine a connu un essor important directement corrélé à l'augmentation des volumes de données à traiter, les ressources humaines jusqu'ici employées ne suffisant plus à la tâche. Comme l'indiquent Brunelli *et al.* [1999], Dimitrova *et al.* [2002] ou encore Snoek et Worring [2005], l'enjeu est donc

1. Cette réplique d'anthologie m'a été suggérée par la lecture assidue de la *Rubrique-à-brac* ® de Marcel Gotlib

de suppléer à la main d'œuvre humaine, la plus sûre mais également la plus coûteuse, par des algorithmes automatiques pour les actions les plus simples.

La thèse que nous exposons est focalisée sur la structuration automatique de documents audiovisuels, un aspect particulier de la thématique indexation. Littéralement, on peut définir la structuration comme la capacité à extraire une organisation interne des documents et contenus à analyser. Il s'agit généralement de dégager du flux audiovisuel des sections véhiculant une information propre et ce de facon automatique, c'est-à-dire en au moyen de méthodes logicielles.

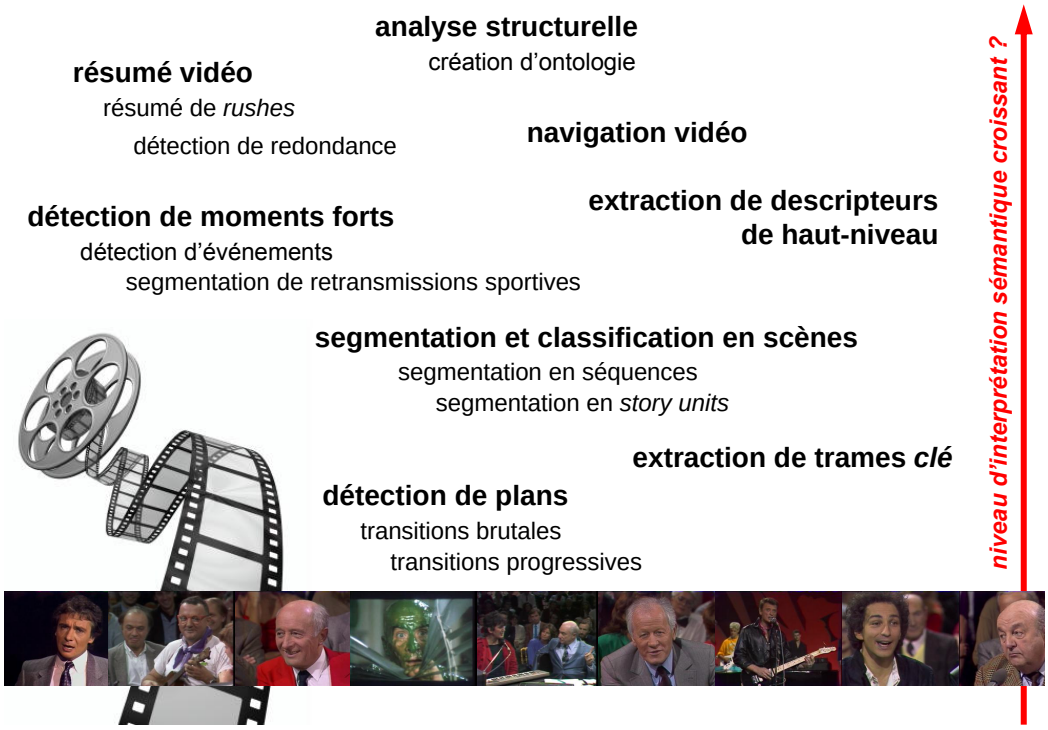


Figure 1 – Terminologie de la structuration automatique.

Cependant, la notion de structuration reste très vague et il n'existe pas à proprement parler de définition consensuelle plus précise. De notre point de vue, la structuration se rapporte à toute tentative d'ordonnancement de contenu. Ainsi, des tâches aussi disparates que la détection de changement de plans de montage, la segmentation en scènes, le résumé automatique, etc. peuvent être considérées comme des processus de structuration. L'approche que nous proposons est orientée vers une structuration de niveau « sémantique » élevé. En effet, le problème fondamental de la structuration est l'écart observé entre l'information extraite automatiquement des données du flux audiovisuel et l'interprétation humaine faite par l'utilisateur de ces mêmes données.

Le problème de la structuration est souvent présenté sous l'angle du désormais célèbre « fossé sémantique » (*semantic gap*). Cette expression est censée incarner l'écart qui s'observe entre

des « concepts » audiovisuels (autre terme dont l'utilisation est parfois hasardeuse) créés par et pour les utilisateurs et leurs représentations numériques directement interprétables en langage machine. Dans cette représentation dichotomique, on accentue les distinctions : détecteurs de bas-niveau/« concepts » de haut-niveau, méthodes numériques/méthodes symboliques, etc. Les auteurs procèdent alors généralement à une hiérarchisation des techniques de structuration en fonction de leur niveau d'interprétation « sémantique » qui, en plus d'être délicate, est sujette à débat (voir figure 1).

Dans le travail présenté par la suite nous préférons une approche plus pragmatique calquée sur celle proposée par Carrive [2007] et identifiant les trois grands acteurs du processus de structuration : les technologies mises en œuvres, les utilisateurs impliqués et les corpus étudiés. De plus, estimant, comme explicité plus haut, que toute tentative d'ordonnancement peut être assimilée à un mécanisme de structuration, nous mettons de côté toute ambition d'organiser de façon « sémantiquement » hiérarchique les tâches de structuration. Nous décidons plutôt de mettre l'accent sur les aspects de granularité et de difficulté des techniques aboutissant à l'élaboration d'une structure automatique.

Contexte d'étude

Né en 1974, à la suite de l'éclatement de l'Office de Radiodiffusion Télévision Française (ORTF), l'Institut National de l'Audiovisuel [2] (Ina) s'est vu assigné comme principales missions d'archiver et de partager toutes les productions radiophoniques et télévisuelles françaises. Il s'agit donc de sélectionner et documenter une partie de ce flux en suivant un certain nombre de règles définies par l'institution. Le renseignement des médias est réalisé par des documentalistes et concerne le catalogage et l'indexation des documents afin de faciliter leur accès. Les volumes gérés par l'Ina soulèvent de nombreux problèmes. Au delà de la veille technologique relative aux difficultés de captation du flux et du stockage de contenus, de nouvelles thématiques de recherche liées à des champs disciplinaires très éloignés ont émergé : sociologie, traitement de signal, informatique, représentation des connaissances, etc. Ainsi, l'Ina peut être considéré comme une plateforme d'exploitation unique en son genre au regard des contraintes volumiques et organisationnelles qui lui sont inhérentes.

Le travail présenté ici a été réalisé dans le cadre d'un dispositif CIFRE (Convention Industrielle de Formation par la REcherche) établi entre l'Institut National de l'Audiovisuel et Télécom ParisTech [3], et plus particulièrement au sein des équipes de recherche respectives *Documentation et Indexation* et *Traitement du Signal et des Images* [4] (TSI).

2. Ina - http ://www.ina.fr/
3. Télécom ParisTech - http ://www.telecom-paristech.fr/
4. TSI - http ://www.tsi.telecom-paristech.fr/

Aperçu des travaux existants

De nombreuses recherches se sont déjà penchées sur l'organisation de documents vidéo aux contenus très structurés et/ou reproductibles. Naturel et Gros [2008] ont par exemple proposé diverses approches pour extraire l'agencement interne de flux télévisuels. Pour cela, les auteurs présentent des méthodes de détection de silence, d'images monochromes, etc. permettant une segmentation en ruptures structurelles comme par exemple un passage publicitaire. Un étiquetage automatique des sections et des corrections de classification est ensuite effectué au moyen de guides de programme (EPG, *Electronic Program Guide*).

De façon similaire, Poli [2008] propose une modélisation des grilles de programmes déjà diffusées afin de prédire les prochaines grilles en les confrontant aux guides de programmes. Pour cela, l'auteur propose une extension contextuelle des modèles de Markov, les modèles de Markov cachés contextuels (CHMM). Enfin, Manson et Berrani [2010] effectuent la même tâche de structuration en identifiant les séquences répétées dans un flux audiovisuel (redondances) puis étiquettent les programmes extraits au moyen de métadonnées.

De nombreuses études ont également été menées concernant la structuration de retransmissions sportives : football (Xiong *et al.* [2003] ou Yu *et al.* [2009]), basket-ball (Zhou *et al.* [2000] ou Xu *et al.* [2003]), baseball (Zhang et Chang [2002] ou Guéziec [2002]), tennis (Kijak *et al.* [2006] ou Delakis *et al.* [2008]) et même formule 1 (Petkovic *et al.* [2002]). Dans tous les cas il s'agit de discerner du flux vidéo des actions caractéristiques du sport considéré : *but* ou *faute* pour le football, *panier* pour le basket-ball, *échange* pour le tennis ou encore *lancer* pour le baseball. De nombreuses méthodes ont été proposées pour la détection de ces moments « forts » du match en cours. Beaucoup reposent sur l'utilisation de chaînes de Markov cachées (HMM) comme les travaux de Assfalg *et al.* [2002] ou Baillie et Jose [2004], ou de réseaux bayésiens dynamiques (DBN) comme les travaux de Delakis [2006] ou Oliver et Horvitz [2005].

Des propositions de structuration ont également été faites pour d'autres types émissions de télévision. On peut ainsi évoquer entre autres les journaux télévisés (projet Infom@gic [5]) pour lesquels une distinction est effectuée entre les parties reportages et celles pendant lesquels le présentateur introduit les sujets. Certains auteurs (par exemple Ide *et al.* [2004] ou Law-To *et al.* [2010]) proposent de retrouver les événements du sommaire introductif et de créer ainsi automatiquement une table des matières du journal.

Enfin, d'autres travaux ont également eu pour objet les débats (Dielmann [2010]), les jeux télévisés (Jaffré [2005]) ou encore les séries et sitcoms (Sivic *et al.* [2009]). Dans ce dernier cas, certains auteurs, comme Cour *et al.* [2008] proposent une identification automatique des personnages par l'analyse des sous-titres.

5. Infom@gic - http ://www.capdigital.com/projet-infomagic-1-2/

Problématique

Dans cette thèse, nous nous intéressons à un type de contenu à notre connaissance peu analysé : le talk show télévisé. En effet, établir une structuration automatique et générique pour ce genre télévisuel comprend à nos yeux un grand nombre de défis. Outre les différences de formats qui peuvent être observées d'une émission à l'autre, les programmes de talk show se distinguent par la variété d'aspects qu'ils peuvent revêtir. Généralement constitués d'alternance de séquence d'interviews, de musique, de reportages, etc., ils sont en effet bien éloignés de programmes dont les montages et réalisations sont beaucoup plus figés et dont l'archétype est le journal télévisé.

De plus, nous souhaitons, à la manière de Khoury [2010] ou Bendris [2011], proposer une organisation complètement non-supervisée et la plus générique possible de ce genre d'émissions. Pour cela nous faisons donc le choix de n'utiliser que les modalités audio et vidéo. Ainsi, nous prenons le parti de ne pas avoir recours à la modalité textuelle et de ne pas adopter de méthodes de transcription automatique de la parole comme par exemple cela peut être proposé dans les travaux de Deléglise *et al.* [2005] ou Guinaudeau *et al.* [2009]. Ce choix est motivé par la volonté de proposer un système aussi universel que possible. En effet, en dépit de l'attrait qu'ils présentent les systèmes de transcription sont nécéssairement dépendants d'une langue et leur utilisation le serait par conséquent également. De plus, ces systèmes nécessitent une adaptation des modèles de transcription pour chaque nouveau corpus traité.

Enfin, un autre aspect d'importance à traiter est celui de la finalité de ce travail : pour qui et pour quoi s'adresse un système de structuration de talk show ? Dans notre cas, nous privilégions une approche multi-scénarii pour laquelle les usagers peuvent à la fois être des professionels de l'archivage ou des utilisateurs issus du grand public. Les utilisations de la structuration de talk show que nous envisageons sont nombreuses et précises. Pour le grand public on peut évoquer la création automatique d'une table des matières de l'émission, l'identification d'événements annotés mais non localisés temporellement, la recherche par recoupement (différentes interprétations d'un même morceau par exemple), etc. D'un point de vue professionel, des outils de navigation et d'annotation semi-automatique peuvent être très précieux pour les compagnies professionelles d'archivages comme l'Ina. Par conséquent, afin de valider la pertinence de la structuration suggérée, nous proposons une expérience utilisateur.

Contributions

L'objet d'étude de cette thèse étant les émissions de talk show, la première étape de notre démarche est d'étudier les spécificités de ce genre télévisuel. En nous inspirant de travaux issus de la communauté des sciences humaines et plus spécifiquement d'études sémiologiques, nous identifions les composantes génériques propres au type d'émissions catégorisées comme talk shows. Nous distinguons en particulier les rôles des invités et présentateurs, le déroulement général du programme, le mode de discours utilisé, etc. Appliquée dans notre cas aux émissions de talk show, cette démarche est généralisable à tout autre type de contenus audiovisuels.

Ensuite, ayant identifié les invariants immuables de ce genre d'émissions et insistant sur le fait qu'un schéma de structuration ne peut avoir de sens que s'il s'inscrit dans une démarche de résolution de cas d'usage, nous proposons une évaluation de l'organisation ainsi dégagée au moyen d'une expérience utilisateur. Celle-ci mesure l'aptitude de sujets à retrouver dans l'émission des événements d'intérêt ayant été préalablement identifiés par les documentalistes de l'Ina avec et sans éléments de structure.

Les études sociologiques et l'expérience utilisateur montrent clairement que la majorité des invariants structurels du talk show peuvent directement être reliés à l'information véhiculée par les prises de paroles des intervenants des émissions. Par conséquent, prenant le contrepied de la grande majorité des travaux issus de la communauté scientifique qui utilisent le plan (*shot*) comme unité fondamentale de structuration, nous proposons d'adopter le tour de parole comme brique élémentaire. Bien que ne présentant pas l'attrait du plan, notamment au niveau de la cohérence visuelle affichée au cours du temps, le tour de parole est porteur de plus d'information « utile ». En effet, le plan est le résultat d'un montage effectué par le réalisateur et ne contient qu'une information « sémantique » partielle sur le déroulement de l'action proposée au spectateur.

Tout naturellement, il découle du choix de faire du tour de parole l'entité atomique de notre approche de structuration la nécessité d'utiliser des méthodes efficaces pour leur détection. Afin de garantir la généricité de la méthode et n'être pas contraint par la dépendance à une langue donnée, les systèmes de transcription automatique de la parole ne sont pas utilisés dans nos travaux. Nous soulignons par conséquent l'importance de la reconnaissance de locuteurs (*speaker diarization*) et lui consacrons la seconde partie de cette thèse. En effet, l'utilisation de méthodes état de l'art de reconnaissance de locuteurs donne des résultats qui, s'ils peuvent être considérés comme satisfaisants au niveau des métriques d'évaluation utilisées, sont difficilement exploitable dans la perspective de structuration que nous proposons. En effet, les corpus d'émissions de talk show présentent des caractéristiques particulières concernant la répartition de la parole. Par conséquent, avançant l'idée que tous les intervenants d'un plateau télévisé sont potentiellement d'importance pour extraire une organisation de ce genre d'émission, nous proposons l'utilisation de deux mesures alternatives au taux d'erreur de reconnaissance traditionnellement utilisé dans la communauté de traitement de la parole.

Nous présentons ensuite les résultats d'un premier travail, réalisé en collaboration avec Simon Bozonnet et Nicholas Evans (EURECOM [6]) et pour lequel un module d'initialisation vidéo basé sur la couleur dominante du costume est ajouté à un système de reconnaissance de locuteurs état de l'art. Ce dernier système, traditionnellement utilisé dans le cadre de captation audio de réunions de travail (*meeting data corpus*), est ici testé sur des émissions de talk show du corpus *Le Grand Échiquier*.

À la vue des résultats offerts par cette première tentative, nous concentrons nos efforts sur la réalisation d'un système original exploitant de manière plus efficace l'information visuelle pour l'apprentissage de modèles de locuteurs (voir figure 2). Ce travail diffère notablement des algorithmes proposés dans la communauté scientifique en suggérant d'utiliser une succession de

6. EURECOM - http ://www.eurecom.fr/

deux regroupements hiérarchiques sur une sélection de plans pour lesquels la personne à l'écran est supposée être la personne qui parle. Le premier de ces regroupements est un *clustering* k-moyennes sous-optimal pour lequel les plans de contenus visuels similaires sont aggrégés alors que le second exploite les caractéristiques audio des partitions nouvellement créées pour effectuer une nouvelle aggrégation par calcul de distances probabilistes en RKHS (espace de Hilbert à noyau reproduisant).

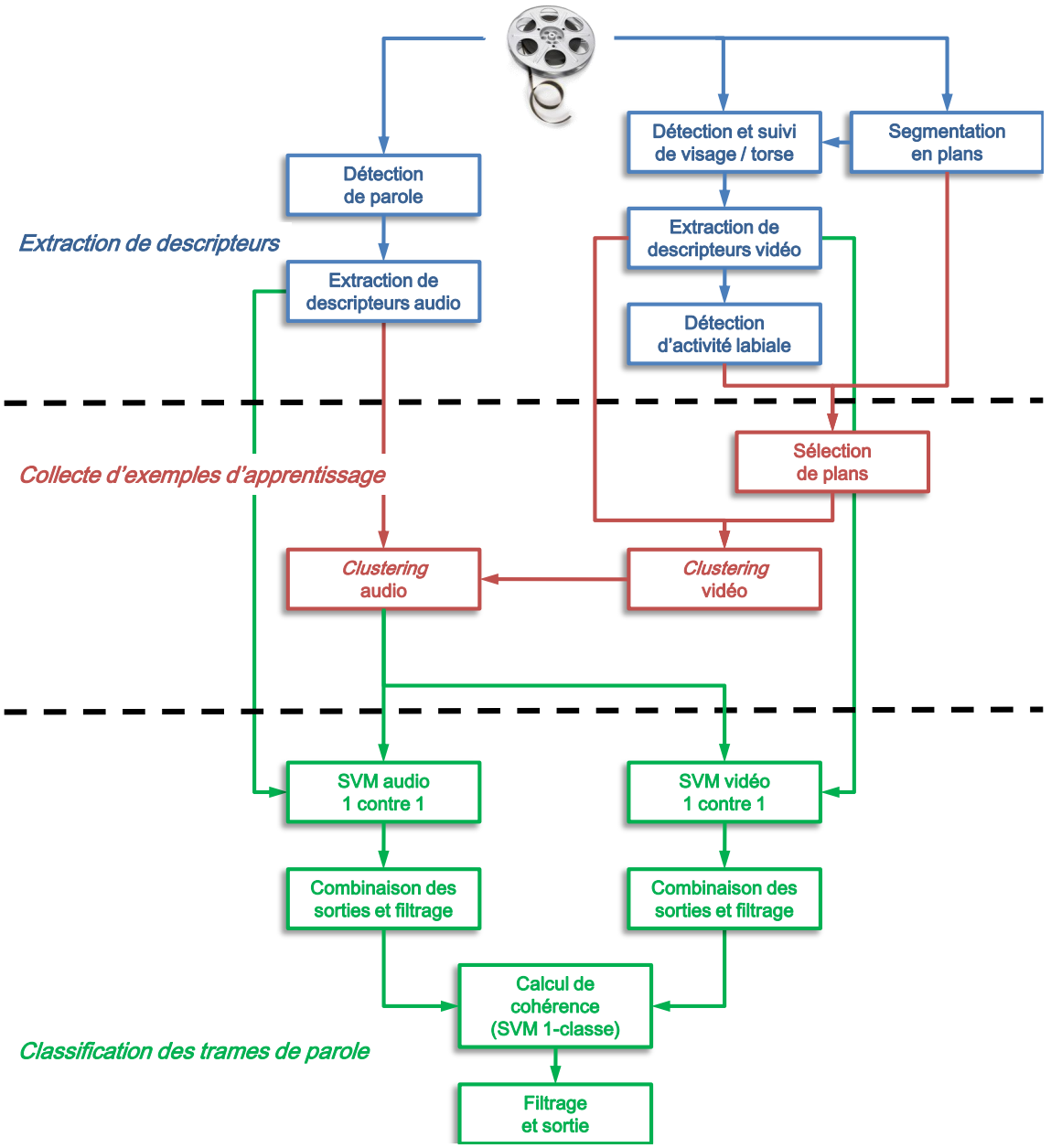


Figure 2 – Présentation de l'algorithme de reconnaissance de locuteurs.

Puis, ayant ainsi collecté de manière non-supervisée des données d'apprentissage fiables des locuteurs présents sur le plateau de l'émission, nous procédons à une classification audio par machines à vecteurs de support (SVM). Nous proposons ensuite l'adjonction à cette architecture d'un module de correction d'erreur de classification par calcul de cohérence entre les sorties SVM audio et vidéo. Cette mesure, effectuée à l'aide de classifieurs SVM à une classe, permet de renforcer la prise de décision en indiquant pour chaque trame de parole laquelle des deux decisions, celle proposée par le classifieur audio ou celle proposée par le classifieur vidéo, est à conserver. Au regard des résultats sur les corpus d'émissions de talk show *Le Grand Échiquier* et *On n'a pas tout dit*, notre système se démarque avantageusement des travaux de l'état de l'art. Il conforte l'idée que les caractéristiques visuelles peuvent être d'un grand intérêt — même pour la résolution de tâches a priori exclusivement audio comme la reconnaissance de locuteurs — et que l'utilisation de méthodes à noyau dans un contexte multimodal peut s'avérer très performante.

En résumé, nos principales contributions sont les suivantes :

- une étude sémiologique du talk show
- une proposition de structure générique du talk show
- une expérience utilisateur validant la structure proposée
- une étape d'initialisation audiovisuelle d'un système de reconnaissance de locuteur état de l'art
- deux mesures de taux d'erreur adaptées aux émissions de talk show
- une nouvelle architecture de reconnaissance multimodale des locuteurs basée sur l'utilisation de méthodes à noyau

Plan du document

Le corps du document est organisé en deux parties de trois chapitres chacune :

La première partie consiste en une proposition de structure pour l'organisation de talk show. Elle débute au chapitre 1 par une analyse du genre télévisuel qu'est le talk show basée sur des travaux issus du domaine des sciences humaines et une comparaison de deux corpus : *Le Grand Échiquier* et *On n'a pas tout dit*. Puis au chapitre 2, après avoir exposé les grands invariants du talk show, nous présentons les résulats obtenus lors d'une expérience utilisateur pour l'évaluation de l'organisation générique que nous proposons. Enfin, nous concluons cette partie en exposant au chapitre 3 des méthodes issues de l'état de l'art ainsi que quelque travaux exploratoires visant à détecter des « concepts » caractéristiques.

À la suite de la première partie, nous dressons le constat que les locuteurs jouent un rôle clé dans le processus de structuration des émissions de type talk show. Fort de cette assertion nous proposons dans la seconde partie de nous focaliser sur la tâche de reconnaissance automatique de locuteurs (*speaker diarization*). Un tour d'horizon des techniques état de l'art utilisées dans ce domaine de recherche et en particulier des méthodes non-supervisées est proposé au chapitre 4. Nous intoduisons également deux mesures d'évaluation de la qualité des algorithmes

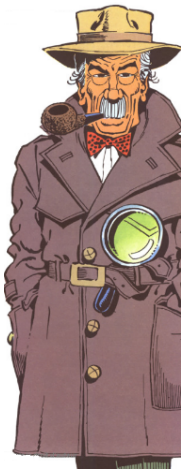
complétant celles traditionnellement utilisées et qui peuvent sembler réductrices dans une persepective de structuration automatique. Un premier système de reconnaisance de locuteur effectué avec Simon Bozonnet et Nicholas Evans (EURECOM) est ensuite présenté au chapitre 5. Enfin, le chapitre 6 est l'occasion de détailler un système de reconnaissance de locuteurs original exploitant de manière plus efficace l'information visuelle et mettant en évidence les bonnes performances pouvant être obtenues avec des méthodes à noyau pour ce type de tâche.

Première partie

Proposition de structure d'émissions de talk show

Nous avons souligné en introduction que chaque processus de traitement vidéo — qu'il s'agisse de techniques de segmentation, classification, détection, etc. — peut être apparenté à une tâche d'organisation du contenu. En effet, chaque tentative d'ordonnancement, à quelque niveau que cela se produise, peut être considérée comme une acception très générale du terme « structuration ».

Par conséquent, travaillant sur les émissions de talk show, nous proposons une étude approfondie de ce genre télévisuel pour en dégager les caractéristiques principales. Puis, une fois ces caractéristiques identifiées, nous évaluons leur pertinence pour la résolution de cas d'usage à caractère industriel au moyen d'un test utilisateur. Enfin, nous présentons des techniques issues de l'état de l'art pour détecter de telles caractéristiques et être ainsi en mesure de proposer l'extraction automatique d'une organisation interne de talk show.



Chapitre 1

Étude du talk show

Introduction

La structuration audiovisuelle revêt un grand nombre de formes, comme explicité dans le chapitre introductif. Cependant, afin qu'une telle structure puisse être considérée d'intérêt pour des applications industrielles mettant en jeu des utilisateurs, il est important de caractériser le genre télévisuel que nous souhaitons traiter dans cette thèse, à savoir les programmes de talk show. C'est pourquoi nous proposons une étude sémiologique pour en identifier les grandes caractéristiques. Une étude comparative entre deux types de talk show est ensuite menée afin de mettre en évidence différences et invariants et de les confronter aux conclusions des travaux de sciences humaines.

1.1 Qu'est-ce qu'un talk show ?

Le talk show est une forme centrale du monde audiovisuel. Il met en présence plusieurs invités autour d'un présentateur pour traiter d'un ou plusieurs thèmes. Les invités sont généralement convoqués pour des raisons précises d'identité en rapport avec le thème traité et sont connus ou inconnus du public. Dans ce dernier cas, il s'agit souvent d'experts ou de représentants d'institutions. Enfin, le présentateur représente l'instance médiatique (voir Charaudeau [1997]) et s'assure du bon déroulement de l'émission.

1.1.1 Histoire des émissions de divertissement

Après la seconde guerre mondiale, les missions de la télévision ont été tacitement définies par les hommes politiques et les intellectuels comme une trilogie. Ainsi, et ce dans tous les pays, on considère que l'audiovisuel doit informer, cultiver et divertir (voir Bourdon [1988]). À cette période, télévisions et radios sont diffusées par des agences nationales comme l'ORTF (Office de Radiodiffusion Télévision Française) en France. Cependant, avec la multiplication des postes de télévision dans les foyers et l'émergence des chaînes privées, le monde de l'audiovisuel connaît une mutation lente mais constante. En particulier, en raison des contraintes de profits, propres à toute industrie et des batailles pour le taux d'audience, le divertissement, parce qu'il est « vendeur », prend de plus en plus de place dans les grilles de programmation. Par conséquent, les

programmes de type jeux télévisés, talk shows, émissions de variétés et télé-réalité occupent une place toujours plus importante du temps d'antenne disponible.

Figure 1.1 – Exemples de talk shows. (A) Cassius Clay et Liberace dans *The Jack Paar Show* (NBC) 1963. (B) Muhammad Ali dans *Apostrophes* présenté par Bernard Pivot (Antenne 2) 1976. (C) Muhammad Ali, Sugar Ray Leonard et Mike Tyson dans *The Arsenio Hall Show* (Paramount Television) 1989.

Certains présentateurs sont devenus célèbres en réussissant à imprimer de leur marque les émissions de divertissement et à faire évoluer le genre. On peut citer par exemple Jack Paar, Johnny Carson, Dick Cavett, Phil Donahue, David Letterman et Oprah Winfrey aux États-Unis ou Guy Lux, Jacques Martin, Jacques Chancel, Bernard Pivot, Michel Drucker et Thierry Ardisson en France. De plus, depuis les années 1990, ces émissions sont devenues un atout considérable pour les candidats politiques. L'ancien président des États-Unis Bill Clinton fut ainsi surnommé le « talk show president » pour ses apparitions remarquées dans *The Phil Donahue Show*, *The Arsenio Hall Show* ou sur MTV.

1.1.2 Une approche sémiologique du talk show

Apparus dans les années 1950, les talk shows télévisés ont connu depuis un succès croissant. De ce fait, les universitaires se sont de plus en plus intéressés à étudier ce type de document et par là-même l'image de la société qu'ils renvoient.
Au regard des taxonomies proposées dans les travaux sémiotiques de Williams [1974], Bourdon [1988] et Charaudeau [1997], les talk shows se positionnent comme des programmes à la croisée des axes informatif, culturel et divertissant. Ceux-ci consistent en général en la présence combinée d'un ou plusieurs présentateurs et d'invités évoquant leur vie professionnelle et/ou personnelle. Il peut s'agir de personnalités connues promouvant leur dernier film, album, émission télévisée, etc. ou bien de personnes inconnues (experts, représentants d'institution, etc.).

Il est cependant difficile de définir le talk show comme un genre. En effet, celui-ci englobe des programmes d'aspects très divers. Comme cela est souligné par Bourdon [1988] et Lochard [1990], la difficulté tient donc à identifier les composantes principales d'émissions dont le contenu et l'apparence peuvent sensiblement changer. Par exemple, les décors de studio et les dispositions de plateau varient considérablement d'une émission à l'autre. De plus, depuis sa création,

le talk show a été adopté dans tous les pays du monde. Par conséquent, des différences culturelles peuvent également être observées. Dans les talk shows français, le public entoure généralement le couple présentateur(s)/invité(s) qui constitue le centre de l'attention alors qu'aux États-Unis, un décor est généralement placé derrière ceux-ci. De plus, le présentateur peut être assis derrière un bureau comme cela arrive souvent dans les émissions américaines ou autour d'une table, avec les invités, comme cela se fait plus couramment en France. Les invités peuvent arriver un par un, tous ensemble, par petits groupes, etc.

Figure 1.2 – Différents plateaux de talk shows.

Dans les talk shows américains, le présentateur est souvent un humoriste (*stand-up comedian*) qui introduit l'émission par un monologue alors qu'en France, celui-ci se contente en général de prononcer une courte présentation du ou des invités et de donner aux téléspectateurs un aperçu des sujets qui seront abordés au cours de l'émission. Enfin, aux États-Unis, une distinction est faite entre le *daytime talk show* qui traite en général plutôt d'affaires publiques, de santé, etc. en interviewant des experts et le plus traditionnel *late-night talk show* qui propose des interludes comiques ou musicaux en plus des discussions menées autour du couple présentateur/invité(s).

Cependant, avec le mouvement général de globalisation qui est observé de nos jours, aussi bien dans le monde audiovisuel que dans les autres industries culturelles, ces différences tendent à s'estomper. Lhérault et Neveu [2003], Lochard [1990] et Munson [1993] ont travaillé à identifier les caractéristiques communes des talk shows, et ce malgré les différences observées. Le plus important des éléments constitutifs de tout talk show est la présence indispensable d'un ou plusieurs présentateurs et d'un ou plusieurs invités. Il n'y a pas de règles quant au nombre exact de participants ; en revanche le couple présentateur/invité est nécessairement le centre de l'attention. La polymorphie de l'émission est également une grande caractéristique des programmes de talk show. En effet, ceux-ci sont généralement construits comme une succession d'interviews, d'interludes musicaux, de reportage télévisuels, d'extraits de film, de jingles, etc.

Les programmes de talk show présentent la particularité d'être structurés autour d'apparentes conversations naturelles. Cependant, comme le montrent Bourdieu [1996], Ghiglione et Charaudeau [1997], Goffman [1981] et Timberg [2002] ces interventions n'en sont pas pour autant moins préparées et formatées. Le présentateur, qui représente l'instance médiatique (voir Charaudeau [1997]), joue un rôle de gestionnaire de la parole. Il questionne, distribue les prises de parole, donne le ton général, oriente le discours, tente d'atténuer les échanges trop vifs, demande des explications et cherche parfois à provoquer des réactions en forçant le trait dramatique ou émotionnel ou en jouant le confident. De plus, il s'assure du bon déroulement de l'émission et l'oriente dans des directions prédéfinies par ses lancements et prises de parole. Chalvon-Demersay et Pasquier [1990] et Penz [1996] ont ainsi montré que le langage utilisé lors de telles émissions est très précisément défini. Il existe par exemple des codes très spécifiques pour introduire un nouvel invité, lancer un interlude musical ou tout simplement changer de sujet de conversation. Par exemple, Chalvon-Demersay et Pasquier [1990] ont étudié la réthorique de l'annonce faite par le présentateur. Pour présenter un invité, il juxtapose les propositions : « C'est une jeune fille, *qui* est très jeune, *qui* est très douée, *qui* est là, *qui* va chanter dans un instant une chanson *qui* j'espère vous plaira. ». Il donne ensuite des précisions promotionnelles : date et lieu de l'exhibition sur scène, titre de l'album, titre de la chanson. Puis, enchaîne avec le dénouement : « Voici... Vanessa Paradis. Vanessa Paradis ! » avec la première nomination, qui clôt le suspense et la seconde, qui ouvre la chanson (Michel Drucker dans l'émission *Champs-Élysées*). Le but est de mettre en place pour le téléspectateur tous les artifices de la conversation spontanée tout en suivant rigoureusement le script de l'émission. Par conséquent, alors que le déroulement de l'émission peut parfois sembler cahotique, le présentateur suit un scénario préalablement établi.

1.2 Comparaison de deux talk shows

À la lueur des travaux sémiotiques commentés précédemment, les composantes principales des émissions de talk show peuvent être identifiées. Cependant, il est intéressant de vérifier la pertinence de ces caractéristiques. Pour cela, deux corpus d'émissions françaises sont étudiés : *Le Grand Échiquier* et *On n'a pas tout dit*. En raison de leurs constructions, formats et années de réalisation, les techniques de production ayant évolué, ces deux émissions peuvent être considérées comme représentatives de la catégorie des programmes de talk show.

1.2.1 Présentation des corpus

Nous présentons ici les corpus *Le Grand Échiquier* et *On n'a pas tout dit*. De plus, nous détaillons les annotations correspondantes qui permettront d'identifier les caractéristiques communes et les différences pour ces deux types de programme. Les propriétés physiques des corpus (encodage, fréquence d'échantillonage, etc.) sont décrites dans l'annexe A.

1.2.1.1 *Le Grand Échiquier*

Le Grand Échiquier est une émission de talk show française créée et présentée par Jacques Chancel et diffusée de façon mensuelle sur la deuxième chaîne de l'ORTF de 1972 jusqu'à fin 1974 puis sur Antenne 2 de 1975 jusqu'à 1989 (218 émissions pour une durée totale de 575 heures). Réalisée en direct et diffusée à une heure de grande écoute pour une durée moyenne de deux heures trente minutes, elle est organisée autour d'un invité principal qui peut être une personne physique ou morale (une équipe sportive ou un orchestre par exemple). Elle alterne les performances de l'invité principal, interviews de ce même invité, des interventions de proches, artistes ou non, de petits reportages en extérieur, des discussions informelles, voir intimes, etc. Les invités sont généralement des personnages de forte personnalité tels Léo Ferré, Lino Ventura, Raymond Devos, Michel Sardou mais également des artistes moins connus du grand public tels Angelo Branduardi, Arthur Rubinstein, Maurice Béjart, Jean-Pierre Rampal, etc. (voir figure 1.3).

Figure 1.3 – Extraits des émissions *Le Grand Échiquier* et *On n'a pas tout dit*.

Le corpus d'étude que nous considérons dans cette thèse est constitué d'un sous-ensemble de cinquante-quatre émissions. Un effort a été apporté pour choisir des émissions représentatives du genre télévisuel qu'est le talk show. Ainsi, le contenu comprend en majorité des émissions consacrées à des musiciens mais également des humoristes, sportifs, organisations caritatives, etc. De plus, *Le Grand Échiquier* ayant été à l'écran pendant près de vingt années, les émissions choisies ont été également retenues afin de couvrir toute la période de diffusion. Une liste détaillée de ces émissions est fournie dans l'annexe B.

Les données vidéo précédemment décrites ne constituent qu'une partie du corpus *Le Grand Échiquier*. La seconde partie est composée des notices documentaires associées à chacun des programmes sélectionnés. Ces notices réalisées en direct par des documentalistes de l'Ina pendant la diffusion des programmes donnent de nombreux renseignements : date de la diffusion, liste des intervenants, résumé de l'émission, liste des œuvres, spécification des ayants droit et autres considérations matérielles comme l'identification des bandes originales par exemple. Ces informations sont détaillées plus précisément dans l'annexe C avec un exemple à l'appui. Les notices sont formatées selon un schéma interne défini par l'Ina. Celui-ci permet des exports aux formats *Word*, *Excel* ou *XML*. Certains des champs, comme celui du résumé, sont en texte libre alors que d'autres sont contrôlés ; ceci afin de créer des liens, vers la base matérielle ou par exemple le thésaurus de l'Ina. Ce dernier consiste en une liste de termes reliés entre eux par des dépendances hiérarchiques et ayant trait à des domaines variés. Il peut ainsi servir de point d'entrée pour des recherches documentaires. L'organisation du thésaurus est décrite plus précisément dans l'annexe D.

1.2.1.2 *On n'a pas tout dit*

L'autre émission étudiée est *On n'a pas tout dit*, un programme présenté quotidiennement et en direct par Laurent Ruquier sur France 2 de septembre 2007 à juillet 2008. D'une durée de 50 minutes exactement, cette émission a pour but de commenter de manière tantôt humoristique tantôt sérieuse l'actualité, juste avant le journal télévisé de vingt heures. En plus de l'animateur principal, des chroniqueurs (généralement cinq ou six) sont présents sur le plateau et alternent sketches, imitations, rubriques cinéma, etc. (voir figure 1.3).

Un sous-ensemble de cinq émissions a été utilisé dans les travaux que nous présentons afin de servir de corpus de validation et attester de la généricité des résultats présentés. Contrairement au *Grand Échiquier*, ce corpus contient exclusivement des données vidéo et a été constitué par Meriem Bendris dans le cadre de son travail de thèse (voir Bendris [2011]).

1.2.1.3 Annotations

Afin de mesurer l'exactitude des techniques de détection, classification, segmentation mises en place mais également de mesurer quantitativement les caractéristiques des émissions de talk show, il est nécessaire d'avoir des annotations (*groundtruth*) à notre disposition. Celles-ci sont de deux natures différentes : annotations audiovisuelles et annotations audio.

Annotations *ELAN*

Nous avons procédé à l'annotation de différents événements audiovisuels d'intérêt pour le corpus *Le Grand Échiquier*. Pour ce faire l'outil d'annotation *ELAN*[1] (voir Slöetjes et Wittenburg

1. EUDICO Linguistic Annotator - http ://www.lat-mpi.eu/tools/elan/

[2008]) développé au Max Planck Institute for Psycholinguistics (*Nijmegen, Pays-Bas*) a été utilisé. Un protocole d'annotation a été mis en œuvre et un annotateur a été engagé pour une durée d'un mois afin de répertorier ces événements d'intérêt. L'annexe E apporte des précisions quant à la manière dont ces annotations se sont déroulées.

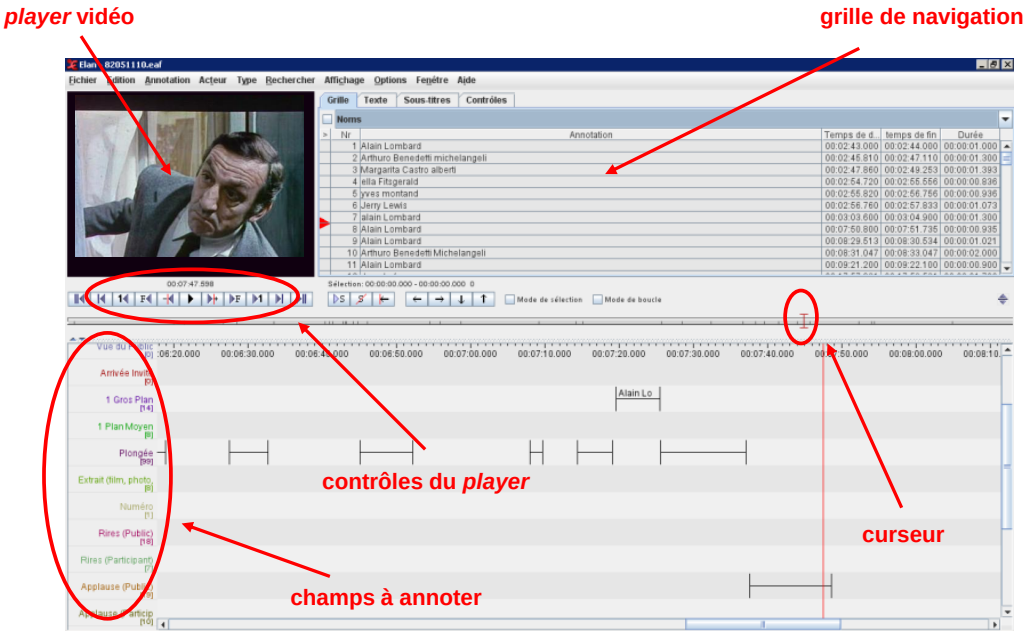


Figure 1.4 – Capture d'écran du logiciel d'annotation *ELAN*.

Au total vingt-trois émissions du corpus *Le Grand Échiquier* ont été annotées en suivant ce protocole. Les événements à répertorier au cours de cette tâche comportent différents niveaux d'information « utile », allant du bas (rires, applaudissements, localisation, etc.) au haut niveau (performances, alignement notice, etc.). Une capture d'écran explicative est donnée dans la figure 1.4.

Annotations *Transcriber*

En plus de l'annotation d'événements audiovisuels d'intérêt nous avons réalisé des annotations précises concernant les segments de parole sur le corpus *Le Grand Échiquier*. Pour cela le logiciel d'annotation audio *Transcriber*[2] développé par Bertin Technologies a été choisi. Il s'agit d'un outil conçu afin de permettre l'analyse de signaux de parole de longue durée. Plus précisément, il est très commode pour les opérations telles que la transcription de parole ou l'identification de tours de parole. Il permet également de créer d'autres méta-données du type condition acoustique, sujet de discussion, applaudissements, musique, etc (voir figure 1.5).

2. Transcriber - http ://trans.sourceforge.net/

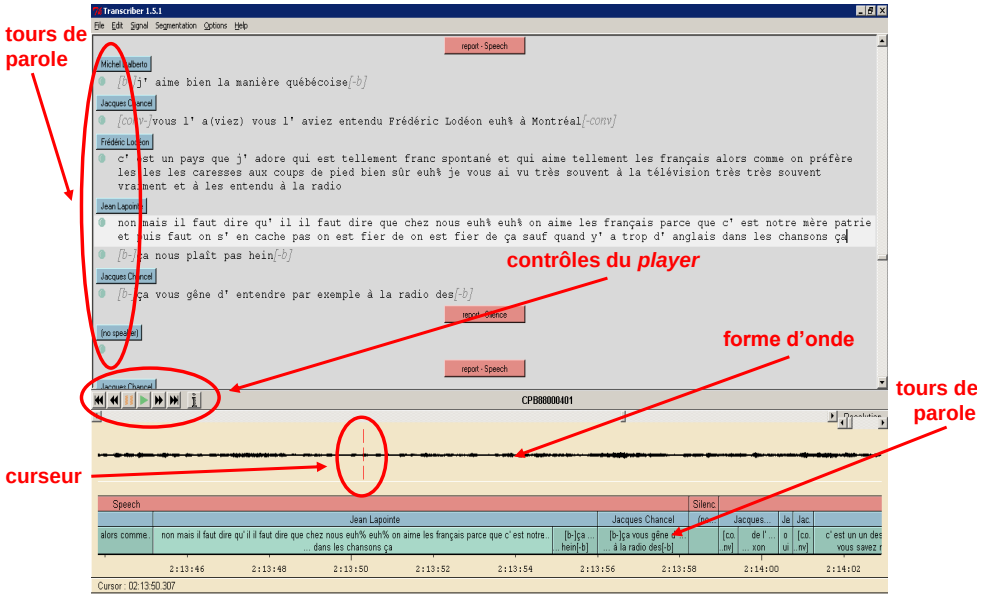


Figure 1.5 – Capture d'écran du logiciel d'annotation *Transcriber*.

Six émissions du corpus *Le Grand Échiquier* ont été annotées en tours de parole, c'est-à-dire en identifiant le locuteur courant. Cela représente plus de quinze heures de programmes constitués environ à moitié de parole. De plus, cinq émissions du corpus *On n'a pas tout dit* ont été annotées par Meriem Bendris dans le cadre de sa thèse (voir Bendris [2011]).

1.2.2 Invariants et différences

Les émissions *Le Grand Échiquier* et *On n'a pas tout dit*, caractéristiques des programmes de talk shows, permettent de mettre en évidence de façon statistique certains attributs communs. Ainsi le tableau 1.1 indique la durée moyenne des éléments constitutifs d'un talk show (« *performances musicales* », « *rires* », etc.). De même, les tableaux 1.2 et 1.3 montrent la répartition des temps de parole pour les différents intervenants ainsi que pour le présentateur. Les résultats proposés ont été obtenus à partir d'annotations manuelles. Pour le tableau 1.1, 22 émissions du *Grand Échiquier* et 5 émissions de *On n'a pas tout dit* ont été utilisées. Cette différence s'explique par des variations beaucoup plus grandes dans le cas du *Grand Échiquier*, le format de l'émission n'étant pas précisément fixé, comme en attestent le tableau 1.1 et les importants écarts types observés. En revanche, pour l'analyse de la parole donnée dans les tableaux 1.2 et 1.3 seules 6 émissions du *Grand Échiquier* ont été annotées. Le corpus *Le Grand Échiquier* a fait l'objet de plusieurs travaux (par exemple Arias *et al.* [2005], Bozonnet *et al.* [2010b], Han *et al.* [2008], Harchaoui *et al.* [2009] et Vallet *et al.* [2010]), de même que *On n'a pas tout dit* (voir Bendris *et al.* [2010b]). Ce dernier corpus est utilisé ici pour valider la généricité de l'approche et peut donc être de plus petite taille.

L'étude des composantes principales des deux talk shows montre des différences majeures. Le tableau 1.1 indique pour chaque émission le nombre et la durée moyenne de quelques événements audiovisuels. Par exemple, les « *performances musicales* » regroupent des moments de l'émission où de la musique est jouée en direct. De façon similaire, les « *performances non-musicales* » concernent les performances proposées au téléspectateur où aucune musique n'est produite. Il peut ainsi s'agir de lecture de poèmes, de numéros de cirque, de scènes de théatre, etc. Enfin, les « *inserts* » correspondent à toutes les séquences diffusées qui ne proviennent pas du montage en direct tels les duplexes, reportages, extraits de film, etc. Certains de ces événements sont non spécifiés, soit parce qu'ils n'existent pas dans le talk show considéré (comme les « *jingles* » pour *Le Grand Échiquier*), soit parce qu'ils ne peuvent pas être calculés. Le nombre de sections de parole est un exemple d'événement difficile à quantifier. En effet, il convient de distinguer les différents mode de discours : lancements, *interviews*, parole sur musique, parole sur inserts, etc. constitutifs de l'entité « *parole* ». Ceux-ci n'ayant pas été annotés, l'estimation du temps de parole a été effectuée par soustraction des différents composants du talk show à la durée totale de chaque émission.

événement	**durée moyenne**		**écart type**		**nombre moyen**		**écart type**	
corpus	GE	OAPTD	GE	OAPTD	GE	OAPTD	GE	OAPTD
émission	165'24" (100%)	50'41" (100%)	25'13" (15.2%)	4'03" (8.0%)	-	-	-	-
participants	-	-	-	-	16.2	7.6	5.7	1.5
perf. musicales	85'43" (51.8%)	36" (1.2%)	25'07" (15.2%)	0 (0%)	26.4	1	8.5	0
perf. non-musicales	9'04" (5.5%)	-	16'17" (9.8%)	-	2.4	-	3.9	-
inserts	16'55" (10.2%)	45" (1.5%)	10'47" (6.5%)	37" (1.2%)	7.8	1	5.8	0.7
parole	40'31" (24.5%)	39'32" (78.0%)	23'31" (14.2%)	4'37" (9.1%)	-	-	-	-
rires	2'35" (1.6%)	3'59" (7.8%)	4'27" (2.7%)	2'35" (5.1%)	41.3	23.3	66.1	10.7
applaudissements	10'34" (6.4%)	5'26" (10.7%)	4'26" (2.7%)	41" (1.3%)	38.2	33.0	18.6	9.1
jingles	-	24" (0.8%)	-	1" (0%)	-	1	-	0

Tableau 1.1 – Statistiques sur les événements audiovisuels pour 22 émissions du *Grand Échiquier* (GE) et 5 émissions de *On n'a pas tout dit* (OAPTD). Les durées sont données en minutes/secondes et les pourcentages reflètent la part de chaque événement dans l'émission.

Comme en témoigne le tableau 1.1, les formats des deux émissions sont sensiblement différents. Dans le cas du *Grand Échiquier*, la durée moyenne du programme est d'environ deux heures trente minutes mais avec de très amples variations alors que pour *On n'a pas tout dit* elle est précisément fixée à 50 minutes. Le motif très reproductible de l'émission *On n'a pas tout dit* peut également être remarqué en observant les faibles écarts types pour les événements « *parole* » et « *applaudissements* », et ce d'autant plus en comparaison avec ceux obtenus pour le *Le Grand Échiquier*. L'absence de « *jingles* » pour *Le Grand Échiquier* peut aussi être notée, ce qui est compréhensible, ce procédé n'étant pas commun à l'époque où furent réalisées ces émissions.

De plus, tandis que *Le Grand Échiquier* présente de grandes variations dans les événements qu'il contient (inserts, passages musicaux, parole, etc.), *On n'a pas tout dit* se compose en vaste majorité de dialogues. En effet, ceux-ci représentent 78% de la durée totale du programme contre 25% ou 39% pour *Le Grand Échiquier* (suivant lequel des tableaux 1.1 ou 1.2 est considéré). Cette différence peut être expliquée par la nature même des deux émissions, centrée sur la vie professionnelle et personnelle d'un ou plusieurs invités dans le cas du *Grand Échiquier*, ou, pour *On*

n'a pas tout dit, sur de petites chroniques, généralement humoristiques comme en attestent les grands nombres d'applaudissements, rires et tours de parole (voir tableau 1.2).

événement	**moyenne**		**écart type**	
corpus	GE	OAPTD	GE	OAPTD
émission	165'17" (100%)	50'41" (100%)	24'41" (14.9%)	4'03" (8.0%)
temps de parole	64'22" (38.9%)	39'32" (78.0%)	21'15" (12.9%)	4'37" (9.1%)
double-voix	5'10" (8.0%)	3'17" (8.3%)	2'31" (3.9%)	1'38" (4.1%)
tours de parole	1264	1348	432	237
1er locuteur dominant	26'08" (40.6%)	14'52" (37.6%)	4'03" (6.3%)	1'49" (4.6%)
2nd locuteur dominant	17'31" (27.2%)	6'15" (15.8%)	6'57" (10.8%)	1'02" (2.6%)
locuteur moyen	4'38" (7.2%)	3'43" (9.4%)	1'02" (1.6%)	1'11" (3.0%)
segment de parole moyen	3.1"	2.2"	0.6"	0.5"

Tableau 1.2 – Statistiques de parole pour 6 émissions du *Grand Échiquier* (GE) et 5 émissions de *On n'a pas tout dit* (OAPTD). Les durées sont données en minutes/secondes et les pourcentages reflètent la part de chaque événement. À l'exception des deux premières lignes, les pourcentages sont calculés sur la durée totale de parole pour une émission.

Un tour de parole, ou tour de locuteur, rend compte de la section de parole pendant laquelle un locuteur B succède à un locuteur A et précède la prise de parole d'un locuteur C (qui peut être à nouveau le locuteur A). Le décompte de ces tours témoigne donc du nombre de changements de parole d'un locuteur à un autre au cours de l'émission. Le nombre important des tours de parole indique des échanges très brefs entre les différents locuteurs présents sur le plateau. Il est intéressant de noter que proportionnellement au temps de parole, beaucoup plus d'interventions sont mesurées dans le corpus *On n'a pas tout dit*. Cela est accentué par l'importance du phénomène de double-voix, ou double-parole qui apparaît lorsque deux locuteurs ou plus parlent en même temps (là encore relativement au temps de parole total). En effet, dans c'est cas pour deux locuteurs A et B intervenant en même temps on compte un tour de parole pour chacun. Les différences observées dans le traitement de la parole en elle-même (comme la longueur moyenne d'un segment de parole) et pour le nombre de tours de parole entre les deux corpus sont typiques des *late-night* et *daytime talk shows* présentés dans la section 1.1.2. Dans *Le Grand Échiquier*, deux personnes partagent environ 70% du temps de parole total - à savoir le présentateur et l'invité principal - alors que dans *On n'a pas tout dit*, la parole est plus « démocratiquement » distribuée comme en témoigne le pourcentage de parole moyen pour un locuteur (voir tableau 1.2). Il est également intéressant de noter qu'en moyenne l'émission *Le Grand Échiquier* comporte neuf participants — au sens d'intervenants et donc de locuteurs — de plus que *On n'a pas tout dit* (16.6 contre 7.6, tableau 1.1). Enfin, les écart-types observés dans le tableau 1.2 indiquent une grande variabilité au niveau de la répartition de la parole. Celle-ci peut s'expliquer par la variété des configurations possibles. Ainsi par exemple, pour un invité principal chanteur ou musicien, on observera généralement un plus grand nombre de « *performances musicales* « et un temps de « *parole* » moins important que pour un humoriste ou un réalisateur (en particulier pour *Le Grand Échiquier*).

Si ces deux émissions peuvent sembler assez radicalement différentes, en particulier en rai-

son de leurs constructions très distinctes, elles n'en ont pas moins des points communs. Il est particulièrement intéressant de noter qu'elles partagent une distribution de la parole très similaire. Dans les deux programmes, la spontaneité du langage est ainsi mise en avant comme en atteste la durée moyenne d'un tour de parole, autour de 2.5 secondes dans les deux cas. De plus, l'importance du phénomène de double-voix renforce ce sentiment de naturel et de familiarité.

événement	**moyenne**		**écart type**	
corpus	GE	OAPTD	GE	OAPTD
temps de parole	23'43" (36.8%)	12'38" (32.0%)	6'58" (10.8%)	1'08" (2.9%)
tour de parole	498 (39.4%)	398 (29.5%)	151 (12.0%)	118 (8.8%)
segment de parole moyen	2.9"	2.0"	0.5"	0.5"

Tableau 1.3 – Statistiques de parole du présentateur pour 6 émissions du *Grand Échiquier* (GE) et 5 émissions de *On n'a pas tout dit* (OAPTD). Les pourcentages sont calculés sur la durée totale de parole de l'émission.

Il est également notable que le présentateur principal (Jacques Chancel ou Laurent Ruquier) est dans chaque émission parmi les deux locuteurs les plus actifs. Cela lui permet ainsi de gérer la distribution de la parole, de questionner ses invités ou d'effectuer des lancements comme cela a été vu dans la section 1.1.2. Il est également intéressant d'étudier ses types d'interventions. Le tableau 1.3 réaffirme l'idée que le présentateur se comporte à la manière d'un chef d'orchestre. En effet, celui-ci intervient très fréquemment tout au long de l'émission comme l'indique son grand nombre de tours de parole. Pour *Le Grand Échiquier*, 39% de ces tours appartiennent à Jacques Chancel alors que pour *On n'a pas tout dit*, 29% sont des interventions de Laurent Ruquier. De plus, il faut noter que ces tours sont en moyenne légèrement plus courts que ceux des autres participants de l'émission. Ainsi, en comparant les tableaux 1.2 et 1.3, on remarque pour *Le Grand Échiquier* que la durée moyenne d'un segment de parole (calculée pour tous les participants, y compris le présentateur) est de 3.1 secondes alors que le présentateur intervient en moyenne 2.9 secondes lorsqu'il prend la parole. De façon similaire, son tour de parole moyen pour *On n'a pas tout dit* est de 2.0 secondes contre 2.2 en moyenne pour tous les locuteurs. Cette différence, même si elle peut sembler minime renforce l'idée d'une séparation bien distincte des rôles dans une émission de talk show. D'un côté le présentateur, de l'autre les invités. Ainsi, aidés par le premier dont le rôle est de les aiguiller, questionner ou faire réagir, les seconds discutent de leurs vie et réalisations professionnelles et/ou personnelles comme cela est détaillé par Chalvon-Demersay et Pasquier [1990] et Penz [1996].

Conclusion

À la lueur des études sémiologiques, nous avons pu identifier des caractéristiques générales aux émissions de talk show. En particulier, nous avons noté l'importance de la parole et de la spontaneité feinte qui l'accompagne. Cette parole est véhiculée par le couple présenta-

teur/invité(s) pivot de ce genre télévisuel. En particulier, le présentateur est responsable de l'articulation logique de l'émission et joue un rôle de chef d'orchestre en assurant les transitions et annonces. L'étude comparative entre les talk shows *Le Grand Échiquier* et *On n'a pas tout dit* confirme l'importance de ces caractéristiques et la necessité d'élaborer une organisation les prenant en compte. Par conséquent, les travaux sémiologiques présentés faisant apparaître l'importance des locuteurs comme éléments de structure permettent de motiver une approche de structuration et serviront de fil conducteur au cours de cette tâche. Il est de plus important de noter que ce type d'approche est généralisable à d'autres contenus audiovisuels. En effet, l'étude de travaux fondés sur l'analyse sémiologique permet la mise en évidence des caractéristiques fondamentales et essentielles d'un genre télévisuel donné.

Chapitre 2

Propositions pour la structuration de talk show

Introduction

L'étude présentée dans le chapitre 1 a clairement mis en évidence certains des éléments caractéristiques des émissions de talk show. Ainsi, l'étude de la distribution du temps de parole indique des rôles d'intervenants très différents, qu'il s'agisse de présentateurs ou d'invités. De plus, les études sémiologiques insistent sur la diversité de « temps » qui est observée pendant de telles émissions, avec des alternances de séquences d'*interviews*, de performances musicales, de reportages, etc. Cependant, l'élaboration d'un système de structuration n'a de sens que si une application concrète peut être dessinée. Nous proposons donc une organisation taxonomique des éléments structurels génériques du talk show présentés dans le chapitre 1 puis, évaluons celle-ci au moyen d'un test utilisateur basé sur un cas d'usage précisément défini : celui de la recherche d'extraits.

2.1 Utilité de la structuration de talk show

Comme explicité dans le chapitre introductif, la structuration audiovisuelle a pour but l'organisation du contenu afin que l'utilisateur ait la capacité d'exploiter celui-ci de la manière la plus optimale possible. Une telle définition est cependant bien trop large pour permettre d'aboutir à la création d'outils génériques. En effet, c'est au regard de la réalisation d'applications très concrètes que peuvent être évaluées les contributions proposées dans le cadre d'un processus de structuration. Se posent alors les questions de pour qui et pour quoi sont créées de telles applications.

Celles-ci se sont formulées dans le cadre de plusieurs projets de recherche, notamment le projet européen K-Space [1], le projet franco-allemand Quaero [2] ou le projet francilien Infom@gic [3] (voir figure 2.1). Il s'agit dès lors de définir des scénarii identifiant les utilisations et les utilisateurs de ces systèmes de structuration vidéo, professionnels (archivistes, documentalistes, etc.)

1. K-Space - http ://www.kspace.qmul.net
2. Quaero - http ://www.quaero.org/
3. Infom@gic - http ://www.capdigital.com/projet-infomagic-1-2/

ou particuliers. Les utilisations peuvent regrouper des applications aussi diverses que le montage assisté, le découpage automatique, l'annotation automatique, la navigation dans des bases de données, la recherche d'extraits, le résumé automatique, la ré-éditorialisation (mise en ligne sur Internet d'un programme déjà diffusé à la télévision), etc.

Figure 2.1 – Logos des projets K-Space, Quaero et Infom@gic.

2.2 Présentation de cas d'usage

Dans le chapitre 1, nous avons mis en évidence un nombre de caractéristiques communes aux émissions de talk show, cela malgré la grande diversité d'apparence de ce type de programme. Ces caractéristiques jettent les bases d'une proposition de structuration. Cependant pour s'assurer de leur pertinence il est utile de se focaliser sur la réalisation de tâches bien définies.

De fait, cette thèse s'étant déroulée dans le cadre d'un dispositif CIFRE (Convention Industrielle de Formation par la REcherche) à l'Institut national de l'audiovisuel (Ina), nous nous positionnons préferentiellement dans un contexte de préservation, d'organisation et de dissémination de fonds documentaires. Dans le cas de l'Ina, il s'agit même de missions assignées par l'État français. Il est possible pour plus de détails de se reporter au document mcc [2008] traitant de la campagne de conservation à long terme des documents d'archives numérisés. On comprend bien que dans un tel contexte, les industries d'archivages sont sensibles aux techniques d'indexation automatique. Il est par conséquent utile de proposer pour ces dernières des cas d'usages pertinents pour évaluer les propriétés d'une bonne méthode de structuration.

Plusieurs cas d'usage nous paraissent pertinents pour évaluer les propriétés d'une bonne méthode de structuration. Ainsi, la sélection d'extraits audiovisuels pour la ré-éditorialisation, c'est-à-dire l'affichage de contenus sur Internet (par exemple pour la création automatique d'une table des matières) peut être avancée. La navigation dans les fonds documentaires pourrait également bénéficier de méthodes d'identification par recoupement, afin de faciliter la recherche inter et intra-document. Par exemple pour rechercher les interventions d'un artiste donné, les différentes versions d'une même chanson, la redondance d'un jingle, etc. Le développement de logiciels facilitant le travail des archivistes lors de l'indexation de documents audiovisuels est une autre possibilité de cas d'usage. En effet, ce travail est jusqu'à présent effectué quasi-exclusivement manuellement. Enfin, l'identification temporelle d'événements d'intérêt, dont la présence est signalée dans les notices documentaires ou les descriptifs textuels joints (souvent au moyen de *tags* ou *labels*), mais pour lesquels les temps de début et fin sont généralement

non-renseignés est un dernier exemple d'application pouvant grandement bénéficier de l' élaboration d'un schéma de structuration.

Des cas présentés, les deux derniers sont particulièrement intéressants puisque de nombreuses applications sont basées sur la recherche d'événements audiovisuels étant donnée leur description. Les compagnies d'archivage comme l'Ina manifestent un intérêt certain pour ce type d'application. Dans le cas de l'Ina, cela peut être d'autant plus intéressant, sachant que pendant très longtemps le contenu des programmes télévisés archivés était annoté sans aucune indication temporelle. En effet, les ordinateurs n'étant apparus pour le grand public que dans les années 1990, les annotations étaient auparavant réalisées en direct par des documentalistes installés devant leur poste de télévision personnel. Par conséquent, une très large partie des annotations audiovisuelles ne possède pas d'indication temporelle telle que la durée d'un événement, son timecode de début, de fin, etc. Un outil facilitant l'alignement temporel de ces notices documentaires serait donc d'un grand intérêt pour l'Ina.

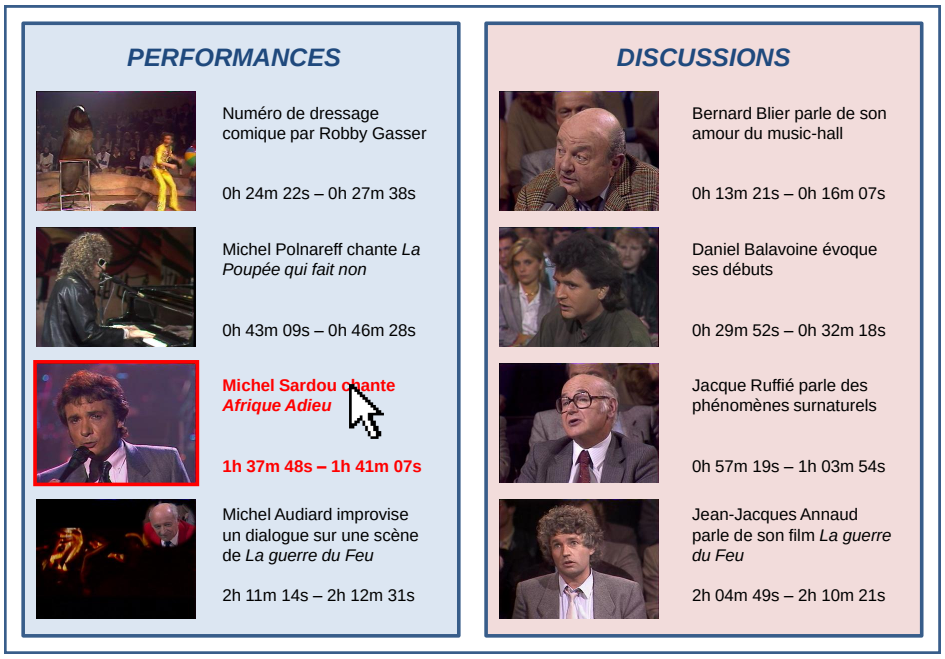


Figure 2.2 – Exemple de visualisation possible pour un cas d'usage *création de table des matières*.

Le problème de structuration de talk show peut alors être défini grossièrement comme l'extraction de composants structurels mis en évidence dans le chapitre 1, leur organisation sous une forme ou une autre pouvant être rattachée à un cas d'usage particulier. L'utilité de ces derniers est de mettre en évidence lesquels de ces grands invariants du talk show sont à prendre en considération pour réaliser un processus de structuration automatique et de proposer ensuite les organisations, visualisations, interfaces, etc. appropriées (voir figure 2.2).

2.3 Composantes génériques du talk show

Les études sémiotiques présentées dans le chapitre 1 et les cas d'usages proposés plus haut nous permettent de mettre en évidence des unités ou composantes de structuration qui sont communes à la grande majorité des émissions de talk show. En effet, la résolution des cas d'usage est amplement facilitée par l'identification de ces invariants structurels du talk show. Nous allons donc les spécifier et évaluer leurs importances respectives en les présentant comme prenant part à trois grandes catégories : *contenu*, *ponctuation* et *localisation*.

2.3.1 Le contenu

Entre les génériques de début et de fin, et ce pendant toute la durée du programme, les éléments de *contenu* se succèdent. Ceux-ci se déclinent en trois grandes entités comme cela est suggéré dans le chapitre 1 : *discussion*, *performance* et *insert*.

Figure 2.3 – Captures d'écran des trois composantes de *contenu* : de haut en bas, *discussion*, *performance* et *insert* pour l'émission CPB84052346 du corpus *Le Grand Échiquier*.

La composante *discussion* englobe tous les moments de l'émission où les participants du talk show — à savoir le présentateur et ses invités — sont dans un acte de conversation. Cette unité structurelle est la plus importante du talk show puisqu'elle assure le liant entre les différentes parties et tient donc lieu d'ossature. Il est bien évident que la composante *discussion* contient la quasi-totalité de l'information « utile » ou « sémantique » disponible. Cependant, comme nous l'avons déjà explicité, nous choisissons de ne considérer cet élément de structure que d'un point de vue « physique » et ainsi de ne pas avoir recours à des technologies d'analyse haut-niveau de type transcription automatique de la parole, ceci afin de conserver une approche de structuration la plus générique possible.

Les *performances* comprennent toutes les actions effectuées sur le plateau et qui ne sont pas de la discussion. Il s'agit en général d'interventions de type artistique. Cela inclut donc les *performances musicales*, mais également les numéros de cirque, les monologues, etc. qui peuvent être catégorisés comme *performances non-musicales* (même si ces derniers peuvent parfois contenir de la musique d'ambiance). Enfin la composante *insert* regroupe toutes les séquences qui sont filmées à l'extérieur du studio d'enregistrement. Il peut s'agir d'images d'archive, de reportages, de vidéos clips, de jingles, etc.

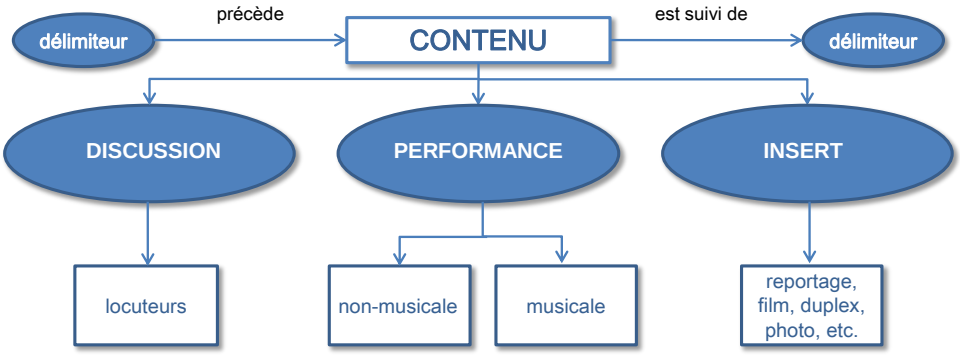


Figure 2.4 – Les différents éléments de *contenu* d'un programme de talk show.

Il est intéressant de noter que ces trois composantes peuvent se recouvrir, partiellement ou complètement. Ainsi, par exemple, les performances commencent parfois avant que le présentateur ait terminé son lancement. On peut demander à un invité de commenter des images d'archive montrées à l'écran, etc. Des exemples variés de ces trois composantes de la catégorie *contenu* sont présentés dans la figure 2.3 et le schéma en blocs 2.4 les résume.

2.3.2 Les délimiteurs

Les éléments de *ponctuation* ou *délimiteurs* servent à assurer les transitions entre les différentes composantes de contenu. Ainsi ces marqueurs fluidifient la succession des différentes parties de l'émission. Ceux-ci peuvent prendre de nombreuses formes. Des événements tels que les *applaudissements* ou les *rires du public* sont considérés comme des *délimiteurs*. De même, du point de vue de la réalisation de l'émission, les *plans de coupe*, les *jingles* et certains *changements de plan* jouent un rôle identique, indiquant une rupture entre deux parties de l'émission bien distinctes. Les *publicités* et autres interprogrammes en sont un autre exemple. Enfin, à un niveau « sémantique » plus élevé, les *transitions* des locuteurs, et du présentateur en particulier, sont d'autres éléments de ponctuation. Cela est confirmé par les études sémiotiques Chalvon-Demersay et Pasquier [1990] et Penz [1996] présentées dans le chapitre 1. De même, ces *transitions* ont été observées en comparant les deux émissions de talk show *Le Grand Échiquier* et *On n'a pas tout dit* dans ce même chapitre. Bien évidemment, dans ces cas, il y a recouvrement avec

la composante de *contenu* → *discussion* (voir figure 2.4). En effet, le présentateur annonce les performances, inserts ou publicités à venir, effectue la transition entre deux sujets de conversation, présente les invités, etc.

Il est très important de noter que les éléments de *ponctuation* donnés par le présentateur ont un rôle essentiel en raison de leur haute valeur « sémantique ». En effet, c'est en général dans ces courtes interruptions que le présentateur donne des indications au spectateur sur ce qu'il va voir et/ou entendre par la suite ou a posteriori ce qu'il a vu et/ou entendu. Les autres éléments de ponctuation présentés servent en général simplement à confirmer et mettre en perspective l'information fournie par les locuteurs et dans la majorité des cas, le présentateur. Le schéma en blocs 2.5 permet une visualisation de ce concept de *délimiteur*.

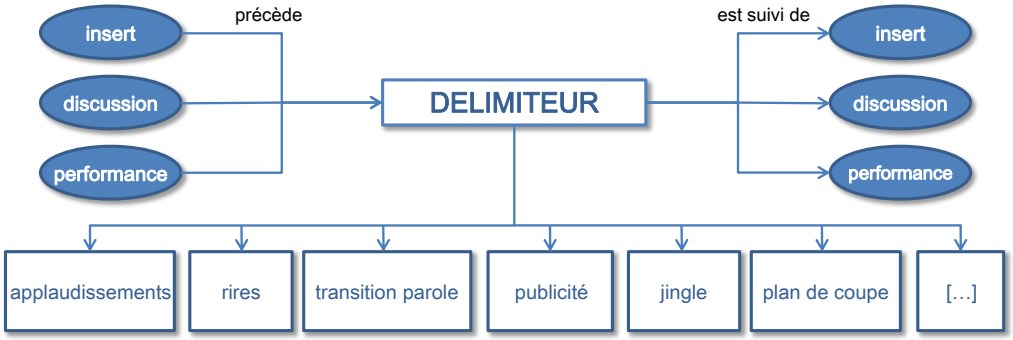


Figure 2.5 – Les différents *délimiteurs* et leur occurence au sein d'un programme de talk show.

2.3.3 La localisation

La dernière grande catégorie de composantes structurelles est la *localisation* de l'action qui est montrée à l'écran. Celle-ci peut se dérouler soit à l'*intérieur* du studio d'enregistrement, soit à l'*extérieur* de celui-ci. Ainsi *extérieur* a trait à toutes les séquences filmées à l'extérieur du studio (rupture spatiale) ou à un instant différé dans le temps (rupture temporelle). Il peut s'agir de duplexes, de reportages, de photos, etc. La composante *intérieur*, elle, reflète la diversité des points de vue filmés dans le studio d'enregistrement et peut prendre plusieurs valeurs telles que *plateau* ou *scène* par exemple. Bien évidemment, ces dernières composantes dépendent fortement du type de talk show traité. Certaines émissions, comme *Le Grand Échiquier*, font ainsi la part belle aux performances artistiques et présentent une unité de structure *scène* importante alors que d'autres, comme *On n'a pas tout dit*, sont essentiellement composées de contenu *discussion* et se déroulent donc pour la quasi-intégralité de l'émission sur le *plateau*. Enfin, il est intéressant de noter qu'un très fort recouvrement est observé entre l'élément de *contenu* → *insert* et celui de *localisation* → *extérieur*.

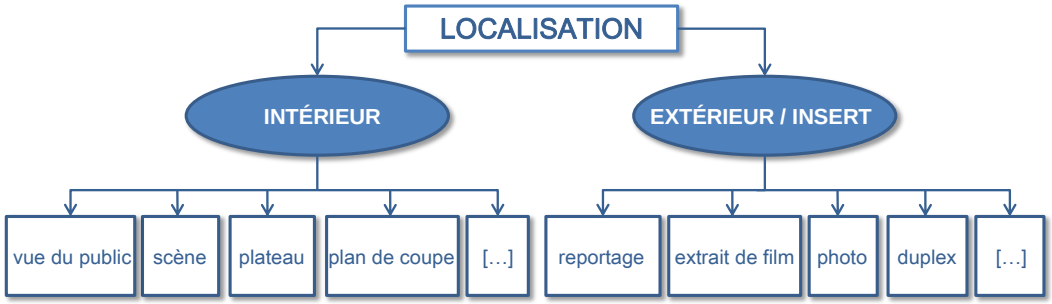


Figure 2.6 – Les différentes composantes de la catégorie *localisation*.

2.3.4 L'importance du locuteur

Il est important de noter que les locuteurs jouent un rôle prépondérant pour établir la structure d'une émission de talk show. En effet, en plus de porter la majeure partie de l'information utile ou humaine, la composante discussion joue un rôle déterminant dans l'enchaînement des différents éléments de contenu. Cela est particulièrement vrai pour le présentateur dont le rôle de chef d'orchestre de l'émission a déjà été souligné. Ainsi, comme prescrit par les règles de montage télévisuel, les performances, extraits, publicités, etc. sont invariablement annoncés. Par ce biais, de précieuses informations sont mises à la disposition du téléspectateur. Suivant le type d'événement introduit, celui-ci se voit donc présenter l'artiste, le titre de la pièce, l'auteur, le compositeur, les musiciens, le réalisateur, etc.

De plus, nous avons insisté sur le fait que nous considérons la composante discussion indépendamment de l'information de haut-niveau qu'elle véhicule, c'est à dire du « sens » des paroles prononcées par les intervenants. Cette prise de position s'explique par le choix de fournir un schéma de structuration indépendant de technologies du type transcription automatique de parole. En effet, de tels systèmes, bien qu'apportant potentiellement de préciseuses informations pour une tâche de structuration, n'en sont pour autant pas suffisamment génériques. Ainsi, en plus d'être délicats à mettre en œuvre, ceux-ci sont dépendants d'une langue donnée. Par conséquent, nous basons notre approche sur l'analyse « physique » de la composante discussion qui peut, notamment grâce à l'étude de la distribution des tours de parole, nous révéler de précieuses informations pour la constitution d'un schéma de structuration.

2.4 Evaluation de la structuration proposée

Les éléments de structure proposés plus haut mettent en évidence le rôle très important du locuteur pour dégager l'organisation interne d'émissions de talk show. Dans le chapitre 3 nous

montrerons comment ces événements peuvent être détectés automatiquement. Cependant, auparavant, afin de s'assurer de la validité du schéma de structuration proposé, nous proposons d'en vérifier l'utilité pour la résolution d'une tâche de structuration. Nous choisissons d'évaluer leur apport à la tâche d'identification temporelle d'extraits audiovisuels. Il s'agit du dernier cas d'usage présenté dans la partie 2.2 pour lequel des descriptions d'événements d'intérêt sont disponibles mais pas leurs localisations temporelles. Dans notre cas, nous proposons de retrouver des événements correspondant à des extraits repérés et documentés par les archivistes de l'Ina dans des émissions de talk show. Il est attendu que les unités de structure présentées dans la partie 2.3 facilitent la navigation des utilisateurs et leurs permettent par conséquent une plus grande efficacité.

Pour procéder à l'évaluation de notre schéma de structuration, nous utilisons le corpus *Le Grand Échiquier* qui a la particularité d'être constitué d'événements audiovisuels de natures très diverses. Ce corpus a également été l'objet de nombreuses études comme celles menées par Arias *et al.* [2005], Bozonnet *et al.* [2010b], Han *et al.* [2008], Harchaoui *et al.* [2009] et Vallet *et al.* [2010]. Les descriptions des événements audiovisuels à retrouver ont été sélectionnées dans les notices documentaires fournies par l'Ina. Un exemple de notice est donné dans l'annexe C).

La tâche proposée aux utilisateurs a consisté à retrouver le plus rapidement possible les événements audiovisuels correspondant à des phrases extraites des notices documentaires. Chaque utilisateur devait retrouver la moitié de ces phrases avec le support des éléments de structuration de la section 2.3 et l'autre moitié sans aucune information. À la fin de l'expérience, les scores de chaque utilisateur ont été calculés comme la somme des temps passés pour retrouver chaque extrait audiovisuel avec et sans éléments de structure. Les phrases issues des notices documentaires ont été sélectionnées manuellement afin de refléter un horizon assez large d'événements audiovisuels de talk show.

2.4.1 Protocole

Les utilisateurs du système ont eu seize extraits audiovisuels à retrouver à partir de quatre émissions différentes, soit quatre extraits par émission. De plus, deux de ces émissions étaient fournies avec les éléments de structure et les deux autres sans. Vingt utilisateurs ont passé le test et le groupe a été divisé en deux sous-groupes : *A* et *B*. Les personnes appartenant au groupe *A* devaient retrouver les extraits audiovisuels avec l'aide des éléments de structure pour les émissions 1 et 3 et sans pour les émissions 2 et 4. À l'inverse celles appartenant au groupe B en bénéficiaient pour les émissions 2 et 4 et pas pour les émissions 1 et 3. De plus, chaque extrait devait être retrouvé en moins de huit minutes. Dans le cas contraire, l'utilisateur se voyait attribué le score maximum de huit minutes.

Afin de garantir de bien mesurer l'apport des éléments de structure dans ce test utilisateur, nous avons utilisé des unités structurelles manuellement annotées. De cette façon, nous nous sommes assurés de ne pas mesurer les performances des détecteurs et ainsi de ne pas mélanger deux effets : l'exactitude des détections des unités de structure et leur utilité pour la résolution de la tâche de structuration considérée.

émission	*Orchestre National de l'Opéra de Paris (1982)*	*Jean-Pierre Rampal (1985)*	*Michel Sardou (1982)*	*Michel Berger (1985)*
durée	2 heures 58 min 37 sec	2 heures 40 min 3 sec	2 heures 51 min 0 sec	2 heures 17 min 10 sec
extrait 1	Ella Fitzgerald chante *Smoke Gets in your Eyes* (insert)	Jean-Pierre Rampal joue le solo de flûte de l'*Orphée* de Gluck	numéro de dressage comique par Robby Gasser et ses otaries	Patrick Vigier présente une guitare à mémoire dotée d'un microprocesseur
extrait 2	l'orchestre dirigé par Alain Lombard joue un *Alléluia*	extraits d'une émission américaine qui montre Rampal dialoguant avec les Muppets	Mireille Darc lit *Colloque Sentimental* de Verlaine	Michel Berger chante *Y'a pas de Honte*
extrait 3	Jacques Chancel s'entretient avec le représentant syndical de l'orchestre	le célèbre flûtiste presente sa famille : Françoise Rampal son épouse, qui fut harpiste, sa fille et son fils	Jean-Jacques Debout, au piano, critique Jack Lang	24 heures de violence : montage de sujets de journaux télévisés A2
extrait 4	mécanicien démontant le piano au studio	Alexandre Lagoya parle de Django Reinhardt	Michel Audiard dialogue une scène de *La Guerre du Feu* de Jean-Jacques Annaud	Daniel Balavoine parle de son 1er album *Mur de Berlin*

Tableau 2.1 – Extraits à retrouver dans les quatre émissions du *Grand Échiquier*.

Le tableau 2.1 et la figure 2.7 présentent les phrases sélectionnées dans les notices documentaires des quatre émissions du *Grand Échiquier* et des captures d'écran des événements audiovisuels correspondants. Idéalement, un outil de navigation dédié à la tâche de recherche des extraits aurait dû être conçu. Cependant, l'évaluation proposée étant axée sur la mesure de la pertinence des éléments de structure et non sur des considérations de présentation, un outil déjà existant a été utilisé. De plus l'évaluation aurait alors mélangé les effets des contributions du nouveau logiciel et des éléments de structure. Par conséquent, les utilisateurs ont effectué la tâche de recherche d'extraits à l'aide du logiciel ELAN (voir Slöetjes et Wittenburg [2008]), un outil professionel pour la création et la visualisation d'annotations audiovisuelles. La présentation, l'interfaçage et l'organisation proposés par le logiciel ont été jugés suffisamment satisfaisants pour considérer les questions d'ergonomie et de maniabilité comme convenablement résolues.

Figure 2.7 – Captures d'écran des 16 extraits à retrouver.

Bien évidemment, quel qu'ait été le scénario dans lequel ils se plaçaient, avec ou sans composantes structurelles, les utilisateurs ont toujours eu à leur disposition les commandes de base de tout *player* vidéo, à savoir : barre de défilement, lecture/pause, déplacement d'une seconde, etc. Afin de se familiariser avec l'outil ELAN, les utilisateurs se sont entraînés pendant quinze minutes avec et sans éléments de structure sur deux autres émissions du corpus *Le Grand Échiquier*. Une fois la phase d'entraînement terminée, la phase de test pouvait démarrer.

Pour chaque émission, les quatre extraits devaient être retrouvés consécutivement et il n'était possible de prendre une pause que lors du passage d'une émission à l'autre. L'objectif était de

limiter le facteur d'« apprentissage ». En effet, les utilisateurs auraient pu, sinon, améliorer leur score en apprenant conjointement le maniement d'ELAN ainsi que l'organisation interne de l'émission sur laquelle ils étaient testés. Une fois un extrait audiovisuel localisé, les utilisateurs devaient entrer dans une interface les temps de début et de fin avant de pouvoir passer à l'extrait suivant.

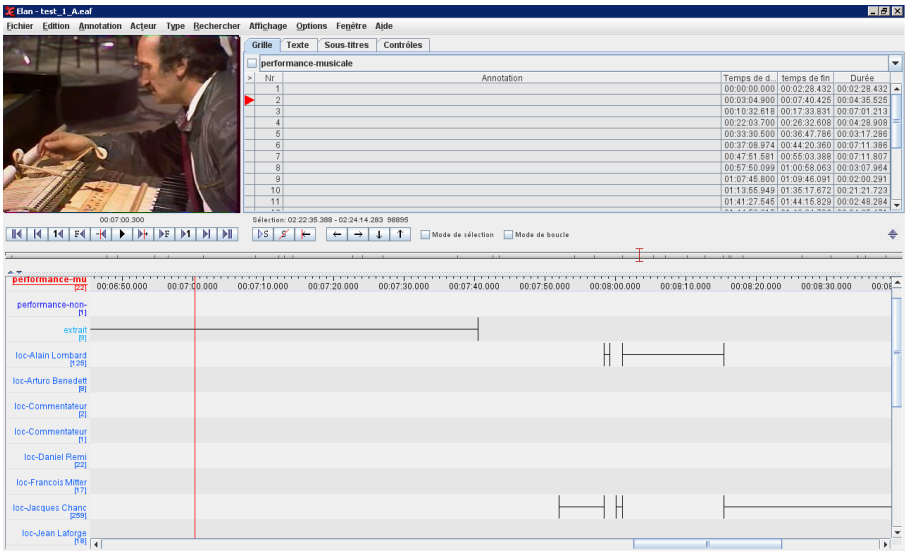


Figure 2.8 – Capture d'écran du logiciel ELAN permettant le parcours des émissions pour la recherche des extraits audiovisuels.

Pour être considéré correctement attribué, un extrait audiovisuel devait présenter un recouvrement d'au moins 70% avec l'annotation manuelle correspondante. Ainsi, l'accent était mis sur la correcte identification des extraits audiovisuels et non la pertinence de leur découpage. En cas de non-détection, c'est à dire une fois la limite de temps atteinte ou en cas de mauvaise attribution, l'utilisateur se voyait octroyer automatiquement le score maximum : huit minutes. L'annexe F présente les instructions telles quelles ont été données aux utilisateurs avec une présentation synthétique du logiciel ELAN. De plus est inclu le questionnaires que ces utilisateurs devaient remplir à l'issue de l'expérience.

contenu				**ponctuation / délimiteur**				**localisation**		
performance		insert	discussion	applaud.	rires	plan 1	plan 2	intérieur		extérieur
musicale	non-musicale		locuteurs					scène	plateau	

Tableau 2.2 – Eléments de structure à la disposition des utilisateurs lors du test.

Le tableau 2.2 récapitule les composantes structurelles mises à la disposition des utilisateurs pour la moitié des émissions. On remarque que celles-ci on été adaptées au format propre du corpus *Le Grand Échiquier* mais que cependant, les grands invariants décrits dans la section 2.3

sont bien présents. Les entrées plan 1 et plan 2 font référence à deux plans de coupe fréquemment utilisés par le réalisateur : une vue du public et une vue en plongée.

2.4.2 Résultats et discussion

Le temps maximum autorisé pour la recherche d'un extrait étant de huit minutes, la durée limite pour le processus d'évaluation était donc théoriquement de 2 h 8 min (sans compter les périodes de pause entre les émissions). Cependant, le temps moyen mis par les utilisateurs était d'un petit peu plus d'une heure : 1 h 1 min 38 s (écart type : 6 min 52 s). Les représentants du groupe *A* ont en moyenne été plus lents que ceux du groupe *B* avec 1 h 4 min 7 s (écart type : 6 min 4 s) contre 59 min 9 s (écart type : 7 min).

extrait	avec éléments de structure				sans éléments de structure			
	temps moyen (sec)	écart type (sec)	% erreur	% temps dépassé	temps moyen (sec)	écart type (sec)	% erreur	% temps dépassé
1	170.7	62.2	0	0	229.1	122.8	0	10
2	438.8	73.7	0	60	442.4	102.1	10	70
3	380.6	131.8	20	30	357.2	108.9	0	30
4	209.1	115.0	0	10	130.3	41.3	0	0
5	392.7	100.9	0	40	387.9	76.4	0	30
6	107.2	39.9	0	0	90.1	34.5	0	0
7	300.2	112.7	0	20	326.5	119.5	10	10
8	115.3	33.3	0	0	327.6	139.5	10	30
9	107.8	87.9	0	0	85.2	22.9	0	0
10	93.5	38.2	0	0	197.7	102.9	0	0
11	164.9	39.6	0	0	241.9	84.1	0	0
12	193.2	44.5	0	0	249.4	105.9	0	0
13	102.3	35.0	0	0	201.9	57.8	0	0
14	319.3	128.0	10	20	236.5	113.6	10	0
15	82.7	36.3	0	0	123.5	63.5	0	0
16	195.8	119.3	20	0	395.4	129.0	20	40
moyenne	210.9	118.4	2.5	11.3	251.4	112.3	3.8	14.8

Tableau 2.3 – Temps mis par les utilisateurs pour retrouver chacun des 16 extraits avec et sans éléments de structure. **% erreur** indique le pourcentage d'utilisateurs ayant entré des temps de début et de fin ne correspondant pas à l'extrait considéré. **% temps dépassé** indique le pourcentage d'utilisateurs qui n'ont pas réussi à localiser l'extrait considéré dans le temps imparti.

Les résultats du tableau 2.3 attestent qu'en moyenne il est plus rapide de localiser un extrait donné avec, que sans éléments de structure. Ainsi, les utilisateurs améliorent leur score de 40 secondes, passant de 4 minutes et 11 secondes à 3 minutes et 31 secondes. Cependant, de façon surprenante, pour certains extraits tels que 3, 4, 6, 9 et 14, l'utilisation des composantes structurelles semble pénaliser les utilisateurs lors de la recherche d'extraits. La taille importante des écarts types doit néanmoins être notée. Pour certains extraits, comme le numéro 4 : « mécanicien démontant le piano au studio », les éléments de structure semblent ne pas avoir notablement aidé les utilisateurs. En effet, aucun de ces éléments n'était clairement lié à la description de l'extrait. Les utilisateurs ont de plus dû s'adapter aux éléments de structure à leur disposition, apprendre lesquels d'entre eux pouvaient s'avérer utiles et dans quelles circonstances. Le fait

d'avoir les composantes structurelles à disposition semble avoir contraint les utilisateurs à s'en servir, et par la même à tenter de qualifier un extrait audiovisuel. *Est-ce un insert ? Une performance ? Cela se produit-il sur le plateau ?* etc. Dans certains cas, l'utilisation des fonctionnalités de base du *player* vidéo s'est avérée plus rapide pour retrouver les extraits. On peut le vérifier en particulier pour ceux facilement identifiables visuellement, comme les extraits 6 (« extraits d'une émission américaine qui montre Rampal dialoguant avec les Muppets ») ou 9 (« numéro de dressage comique par Robby Gasser et ses otaries »). Cela confirme l'idée que les utilisateurs peuvent avoir tendance à « oublier » d'utiliser les fonctionnalités de base du *player* video tels que la barre de défilement, avance rapide, etc. Cette inclinaison paraît cependant facilement rectifiable.

Pour d'autres extraits, comme par exemple le numéro 3 « Jacques Chancel s'entretient avec le représentant syndical de l'orchestre », de nombreux utilisateurs ayant eu à disposition les éléments de structure ont eu tendance à procéder rapidement et à assigner le rôle de « représentant syndical de l'orchestre » à la mauvaise personne. Cela peut s'observer au regard du grand nombre d'erreurs pour cet extrait. Les utilisateurs ayant les composantes structurelles à disposition ont plus souvent mal localisé cet extrait. De plus, puisque les erreurs ont été pénalisées avec le temps maximum autorisé (8 minutes), le temps moyen calculé s'en est trouvé considérablement augmenté. Cela confirme l'idée qu'une fois à leur disposition les utilisateurs ont tendance à faire confiance quasi-aveuglément aux éléments de structure. Ils semblent alors prendre une part moins active dans le processus de recherche et, sentant que les composantes structurelles leurs fournissent des informations supplémentaires et un avantage comparatif, sont alors plus prompts à répondre dans les limites de temps autorisées.

catégorie	**avec éléments de structure**				**sans éléments de structure**			
	temps moyen (sec)	**écart type (sec)**	**% erreur**	**% temps dépassé**	**temps moyen (sec)**	**écart type (sec)**	**% erreur**	**% temps dépassé**
discussion	193.2	101.2	3.8	6.3	287.2	74.2	5	13.8
insert	152.6	55.0	0	2	164.5	70.3	0	2
perf. musicale	285.7	142.4	2	24	277.2	140.1	4	22
perf. non musicale	165.2	62.7	0	3.3	192.5	59.7	0	0

Tableau 2.4 – Temps mis par les utilisateurs pour retrouver chaque catégorie de composantes génériques de structure : *discussion*, *insert*, *performance musicale* et *performance non-musicale*, avec et sans éléments de structure.

Il est possible de classer les 16 extraits selon quatre catégories qui ont été définies comme composantes structurelles génériques des programmes de talk show, à savoir : *discussion*, *performance musicale*, *performance non-musicale* et *insert*. En raison de leur nature, certains extraits peuvent appartenir à plus d'une catégorie. Ainsi par exemple l'extrait 12, « Michel Audiard dialogue une scène de *La Guerre du Feu* de Jean-Jacques Annaud », appartient aux catégories *performances non-musicale*, *discussion* ainsi qu'à *insert*, puisque le film est montré à l'écran.

Le tableau 2.4 donne les temps moyen mis pour retrouver chacune des quatre catégories d'éléments. À l'exception des *performances musicales* pour lesquelles les utilisateurs n'ayant pas eu accès aux éléments de structures ont réussi légèrement plus rapidement à retrouver les ex-

traits, les résultats du tableau 2.4 confirment l'avantage obtenu par les utilisateurs avec les éléments de structure. Cela est particulièrement frappant pour la catégorie *discussion* pour laquelle ces derniers sont beaucoup plus performants. C'est encore confirmé par les faibles pourcentages d'utilisateurs dépassant la limite de temps. Il semble que dans ce dernier cas, l'information soit effectivement plus facile à retrouver.

À la suite du test, les utilisateurs ont dû remplir un questionnaire leur demandant un retour d'expérience. En particulier, ils devaient indiquer quels éléments de structure ils avaient utilisés. Le tableau 2.5 regroupe les réponses données dans ce questionnaire. Comme on pouvait s'y attendre, les éléments de structure de la catégorie *contenu* ont été très appréciés des utilisateurs. Cela s'explique par la nature même des événements audiovisuels à retrouver, ceux-ci étant tirés des notices documentaires. Au contraire, il n'y avait aucun extrait à retrouver du type « *X* rit à une blague de *Y* » ou « *W* applaudit la performance de *Z* », etc. ce qui explique les faibles scores pour les composantes rires, applaudissements, etc. Le plus faible score d'utilité pour les performances non-musicales s'explique par le manque d'événements y correspondant pour les utilisateurs du groupe *B*. En effet, ceux-ci n'avaient pas à leur disposition les éléments de structure pour la troisième émission au cours de laquelle deux performances non-musicales étaient à retrouver (extraits 10 et 12). Enfin, il est intéressant d'observer que, tandis qu'en moyenne il était légèrement plus long de retrouver les performances musicales à l'aide des éléments de structure que sans, cet attribut a été néanmoins considéré comme très utile par les utilisateurs.

élément	perf. musicale	perf. non musicale	insert	locuteur	appl.	rires	plan 1	plan 2	scène	plateau	ext.
utilité	3.9	2.9	3.6	3.9	1.5	1.4	1.4	1.4	1.9	1.9	1.8

Tableau 2.5 – Utilité des différents éléments de structure selon les utilisateurs (4 : très utile, 3 : assez utile, 2 : peu utile, 1 : inutile).

Il est cependant utile de noter que l'importance des écarts types mesurés semble indiquer que le logiciel ELAN ne dispose peut être pas de toutes les qualités escomptées, notamment au niveau de l'ergonomie. Le logiciel joue en effet un rôle essentiel dans la présentation et la navigation parmi les composantes structurelles. Il est donc difficile d'isoler les seuls mérites de celles-ci sans mesurer conjointement plusieurs effets. Ainsi, les résultats dépendent également de la capacité de l'utilisateur à utiliser de façon efficace le logiciel, de sa stratégie de recherche aussi bien que d'un facteur « chance » impossible à écarter ni à quantifier. Tous les utilisateurs étaient familiers des outils informatiques. Cependant, de notables différences ont été observées dans leurs manières de rechercher les extraits. Les utilisateurs les plus rapides avaient tendance à utiliser la barre de défilement pour parcourir l'intégralité de l'émission avant de rechercher les événements demandés et ce, qu'ils aient eu ou non les éléments de structure à leur disposition. Ces sujets se sont donc fait une idée rapide de l'organisation de l'émission. Le temps passé à cette manœuvre a ensuite pu être ratrappé lors de la recherche d'extraits, les utilisateurs étant plus à même d'identifier les parties de l'émissions concernées. Ce facteur d'« apprentissage » est un aspect très important et difficile à isoler. Lorsqu'ils recherchent des extraits, les utilisateurs accumulent expérience et information qui vont parfois leur permettre de retrouver plus rapidement de futurs événements (dont ils n'ont pas connaissance à ce moment). Cela est clairement

lié à la capacité de mémorisation et d'identification de l'utilisateur, mais également au facteur « chance ». De plus, malgré la séquence d'entrainement, les utilisateurs améliorent tout au long de l'émission l'utilisation qu'il font du logiciel ELAN.

2.4.3 Conclusions de l'évaluation

L'évaluation menée ici vise spécifiquement la résolution d'un cas d'usage donné pour attester de l'utilité du schéma de structuration audiovisuelle proposé dans ce chapitre. L'intérêt de l'ensemble des composantes structurelles présentées dans la partie 2.3 tient en la possibilité de les utiliser également pour d'autres tâches comme la rééditorialisation, la recherche par recoupement sur plusieurs émissions, par exemple d'un artiste donné dont on sait qu'il est présent dans les émissions, etc. Dans de telles circonstances, la structuration proposée permettrait également une simplification et une amélioration des performances pour des tâches qui sont d'ordinaire quasi-exclusivement réalisées manuellement. Cette évaluation est également un point de départ intéressant pour mener une réflexion sur la construction d'outils de navigation intra-programme efficaces.

Les résultats de cette expérience utilisateur prouvent l'utilité des éléments génériques de structure proposés dans la section 2.3. Les utilisateurs montrent en effet une plus grande facilité à retrouver les extraits audiovisuels qui leurs sont proposés avec ceux-ci à disposition. Cela est particulièrement observable pour les extraits de type *discussion*. Pour les autres genres d'événements, la différence n'est pas aussi marquée. Par conséquent, l'utilité du schéma de structuration proposé dans ce chapitre et basé sur l'étude de travaux sémiotiques a été confirmée par ce test utilisateur. Cependant, la navigation au sein des émissions, réalisée dans l'expérience avec le logiciel ELAN, semble perfectible, indiquant l'importance de la prise en compte des questions d'ergonomie.

Conclusion

Nous avons mis en évidence dans ce chapitre l'importance de définir clairement le cadre dans lequel s'inscrit tout mécanisme de structuration. En effet, la validité d'une organisation automatique de document sera toujours tributaire de l'application visée. Cependant, nous avons pu noter que les grands invariants du talk show mis en évidence dans le chapitre 1 restent toutefois valides pour la grande majorité des applications. Nous avons ensuite défini plus précisément ces invariants structurels, constitutifs de n'importe quelle émission de talk show et qui se regroupent en trois grandes familles : « contenu », « délimiteurs » et « localisation ». Le mécanisme de structuration proposé étant constitué de l'ensemble de ces caractéristiques du talk show, nous avons mesuré le bien fondé d'une telle organisation en mettant en œuvre une évaluation. Au cours de celle-ci des utilisateurs ont eu à retrouver de façon la plus efficace possible des

extraits audiovisuels selectionnés du *Grand Échiquier* dans les notices documentaires associées avec et sans le concours de la structuration proposée. Les résultats ont mis en évidence l'avantage d'un tel outil pour ce cas d'usage et il est prévisible que des résultats similaires auraient été obtenus pour d'autres applications.

Chapitre 3

Détection d'éléments de structure

Introduction

L'évaluation de la structuration proposée dans le chapitre 2 a clairement indiqué l'intérêt de l'organisation que nous proposons pour la résolution de cas d'usage à caractère industriel comme la recherche d'extraits audiovisuels. Cependant, cette évaluation a été réalisée au moyen de données annotées manuellement afin de garantir que les résultats mesurés reflétaient bien uniquement le degré d'utilité des éléments de structure fournis et non pas les performances des détecteurs. Nous présentons par conséquent dans ce chapitre des méthodes issues de travaux de la communauté scientifique pour la détection automatique de ces éléments de structure. Leur utilisation pour une tâche de structuration est motivée par les niveaux de confiance affichés garantissant de bons résultats. Parmis les techniques proposées, nous nous attachons particulièrement à étudier la tâche de reconnaissance de locuteurs. En effet, comme cela a été précédemment mis en évidence, de précieuses informations peuvent être extraites des signaux de parole même si le choix est fait de ne pas avoir recours à des techniques de transcription automatique.

3.1 Liens entre éléments de structure et détecteurs

Ayant proposé dans le précédent chapitre une organisation générique des émissions de talk show basée sur des éléments structurels, nous souhaitons à présent étudier dans quelle mesure ceux-ci peuvent être localisés automatiquement. Pour cela, plusieurs détecteurs utilisant les informations portées par les modalités audio et/ou vidéo peuvent être implémentés. Le tableau 3.1 donne une présentation compacte des éléments génériques de structuration présentés au chapitre 2 et associe des détecteurs en vue de l'identification automatique de chacun de ces composants. Comme nous l'avons précédemment montré, il existe trois grandes familles d'éléments de structure : *contenu*, *ponctuation* (ou *délimiteurs*) et *localisation*. Chacune de celles-ci peut ensuite être subdivisée comme par exemple : *contenu* → *performance* → *non-musicale* qui contient toutes les interventions artistiques de type monologue, lecture, numéro de cirque, scène de théatre, doublage de film, sketch, etc. Comme on peut le voir, la plupart des composantes structurelles proposées peuvent être directement associées à des sorties de détecteurs traitant les flux audio et vidéo. Nous allons présenter dans la suite de ce chapitre les méthodes

de détection qui peuvent être utilisées et montrer que de bons résultats peuvent en être attendus lorsque les horizons de décision sont grands.

éléments structurels			détecteurs associés	exemples de travaux
Contenu	parole	locuteurs	reconnaissance de locuteurs	Fredouille *et al.* [2009]
	performance	musicale	détection de musique	Richard *et al.* [2007]
		non-musicale	reconnaissance de locuteurs détection de rires détection d'applaudissements	Fredouille *et al.* [2009] Wilkins *et al.* [2007] Hory et Christmas [2007]
	inserts	photo	détection d'images fixes	Swain et Ballard [1991]
		film - reportage	détection de sons environnementaux changement dans la distribution de couleur détection de bandes noires	-
Ponctuation	transition - lancement		reconnaissance de locuteurs	Fredouille *et al.* [2009]
	applaudissements		détection d'applaudissements	Hory et Christmas [2007]
	rires		détection de rires	Wilkins *et al.* [2007]
	plan de coupe		détection de plans similaires	Lowe [2004]
	jingle		détection de musique détection de plans similaires	Pinquier et André-Obrecht [2004]
	publicité		changement dans la distribution de couleur détection de trames monochromes détection de changements de plans	Baghdadi *et al.* [2008]
Localisation	intérieur	scène	détection de scène	-
		plateau	détection de plateau	-
	extérieur		changement dans la distribution de couleur detection de sons environnementaux	Swain et Ballard [1991] Wilkins *et al.* [2007]

Tableau 3.1 – Liens entre éléments génériques de structuration et détection automatiques.

3.2 Organisation

Nous proposons dans la suite de ce chapitre une sélection de méthodes issues de l'état de l'art et relativement simples à mettre en œuvre pour la détection d'éléments structurels génériques du talk show. Plutôt que de les présenter de façon itérative en suivant l'ordre du tableau 3.1 nous proposons de regrouper ces invariants audiovisuels des émissions de talk show par approches de détection. Pour cela, malgré les réticences affichées en introduction, nous choisissons une approche pseudo-hiérarchique présentée dans la figure 3.1. Ainsi, nous distinguons les méthodes de segmentation, de détection de « concepts » de haut-niveau et de détection de « concepts » de niveau supérieur.

Les premières sont exclusivement non-supervisées, c'est à dire qu'elles ne nécessitent aucune injection de connaissance et sont de ce fait très génériques. Les secondes en revanche se présentent habituellement sous la forme de problèmes de classification pour lequels des exemples d'apprentissage sont indispensables. Il est question dans ce cas de méthodes supervisées. Enfin, les dernières concernent la détection d'événements plus riches « sémantiquement ». Ces méthodes sont elles aussi supervisées mais utilisent fréquemment des combinaisons de résultats de détections de plus bas niveau. Il est en effet bien souvent compliqué de construire des classifieurs simples pour des événements du type publicité ou performance non-musicale.

Beaucoup des résultats présentés par la suite, sont issus de travaux d'équipes de recherche

participant au projet francilien Infom@gic [1](voir Campedel et Hoogstoël [2011]) qui regroupe des partenaires académiques (Télécom ParisTech, Université Paris 6, etc.) et industriels (Ina, EADS, Pertimm, Xerox, etc.).

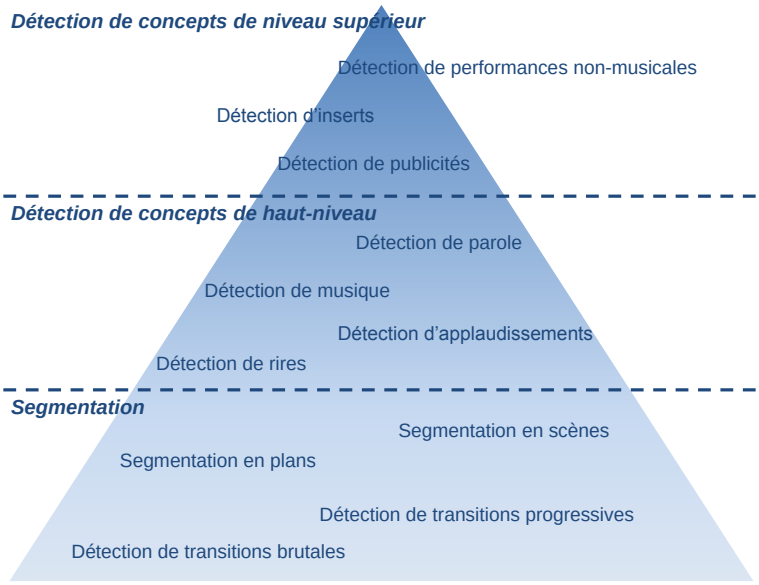


Figure 3.1 – Proposition de hiérarchisation de différentes méthodes de détection d'événements audiovisuels.

3.3 La segmentation

On peut généralement présenter les problèmes de segmentation comme le découpage d'un flux continu en régions « homogènes ». Ce flux peut être audio, vidéo ou audio et vidéo. De plus, on définit l'homogénéité en fonction de la tâche à accomplir. Comme nous l'avons précisé précédemment, les méthodes de segmentation ne nécessitent pas l'introduction d'exemple d'apprentissage et sont de ce fait qualifiées de non-supervisées.

3.3.1 La segmentation en plans et en scènes

La segmentation en plans a été pendant longtemps le problème canonique du domaine de l'analyse de contenu audiovisuel. En effet, le plan est considéré dans la majorité des travaux de structuration audiovisuelle comme l'unité fondamentale ou unité atomique au sens de plus petit élément constitutif d'un document audiovisuel. On définit le plan comme une séquence d'images capturées sans interruption par une caméra. Deux types de transitions sont généralement considérées pour passer d'un plan à un autre : la transition brutale sans recouvrement

1. Infom@gic - http ://www.capdigital.com/projet-infomagic-1-2/

d'images (*hard cut*) et la transition progressive (voir figure 3.2). Cette dernière relève d'effets d'éditions et peut être déclinée sous diverses formes. On parle ainsi de fondu à l'ouverture (*fade in*), de fondu à la fermeture (*fade out*), de fondu enchaîné (*dissolve*), de balayage (*wipe* : passage d'une ligne verticale ou horizontale pour faire apparaître un nouveau plan), etc.

(a) Une transition brutale (*cut*)

(b) Une transition progressive : fondu enchaîné (*dissolve*)

(c) Une transition progressive : balayage (*wipe*)

(d) Une transition progressive : effet fantaisie (*wipe*)

Figure 3.2 – Exemples de transitions entre deux plans (d'après Poli [2007]).

Depuis 2001, le programme d'évaluation TRECVid[2] est devenu une référence en matière d'analyse de contenu vidéo en proposant des bases de données et méthodes d'évaluation aux organismes intéressés à prendre part à de telles campagnes. De nombreuses tâches de structuration vidéo sont proposées annuellement aux institutions participantes : détection d'événements, création de résumés, détection de copies, etc. Bien évidemment, un problème de découpage en plans est également donné (du moins jusqu'en 2007, la tâche étant considérée depuis comme résolue). Smeaton *et al.* [2009] présentent un retour sur toutes les méthodes soumises au cours des sept premières années (de 2001 à 2007) et leurs résultats respectifs.

Pour la détection de transitions brutales parmi les nombreuses techniques proposées comme la comparaison de contours, de vecteurs de mouvement ou de compression, les techniques

2. TREC Video Retrieval Evaluation - http ://trecvid.nist.gov/

les plus simples basées sur la comparaison d'histogrammes de couleur sont les plus performantes. Cela a été mis en évidence par les études comparatives de Lienhart [1998], Brunelli *et al.* [1999] ou encore Browne *et al.* [2000]. Pour les transitions progressives, de nombreuses méthodes existent également. Par exemple, Petersohn [2004] propose une méthode pour la détection de deux types de transitions progressives, à savoir : les fondus (*dissolve*) et les balayages (*wipe*). Chacune de ces transitions est détectée séparément à partir d'une combinaison d'informations sur la couleur, le mouvement et les contours. La précision obtenue avec cette méthode est de 85% pour un rappel de 70%. La méthode utilisée par Volkmer *et al.* [2004] se démarque par l'utilisation d'une fenêtre glissante par comparaison des couleurs sur les parties avant et arrière de la fenêtre. L'avantage de la fenêtre glissante est qu'elle permet de tenir compte des dégradations des images sur un long intervalle temporel, plutôt que trame par trame. Il est cependant important de noter que généralement les transitions brutales représentent la vaste majorité des transitions pour les programmes télévisés. De ce fait, des résultats très satisfaisants peuvent être attendus lors du découpage en plans d'émissions de talk show par les méthodes issues de l'état de l'art.

Le plan est un objet vidéo parfaitement défini et la détermination de ses bornes temporelles est un problème bien posé. Cependant, celui-ci est porteur de relativement peu de « sens ». En effet, l'information « utile » qu'il véhicule est d'une portée assez limitée. Certaines études se sont donc penchées sur la notion de scène audiovisuelle. La définition d'une scène est plus subjective que celle d'un plan. Il s'agit en général de plans adjacents partageant une même unité de sens. De nombreux synonymes au mot anglais *scene* sont employés : *sequence* (Aigrain *et al.* [1997]), *story unit* (Yeung et Yeo [1996]), *logical unit* (Vendrig et Worring [2002]), *video paragraph* (Wactlar *et al.* [1996]) ou encore *act* (Aoki *et al.* [1996]). De toute évidence, il est néanmoins nécessaire d'effectuer une bonne segmentation en plans pour réaliser ensuite un découpage en scènes audiovisuelles satisfaisant.

3.3.2 La segmentation audio

Dans la communauté audio, la segmentation (*audio diarization*) est un problème très classique (voir Reynolds et Torres-Carrasquillo [2005]). De nombreuses tâches peuvent être traitées de façon non-supervisée comme par exemple la segmentation en grandes classes audio (musique, parole, bruit, etc.). Il est alors fréquent de faire référence à ces tâches comme détection de changements.

La segmentation en tours de parole est généralement réalisée de manière non-supervisée. Comme nous le verrons plus en détail dans la seconde partie (chapitre 4) il existe diverses manières de procéder à ce découpage. Il est ainsi possible d'utiliser des méthodes génératives qui tentent de modéliser les données, par exemple au moyen de modèles de mélanges gaussiens (GMM, voir Duda *et al.* [2000]) comme cela est proposé par Chen et Gopalakrishnan [1998] et Gish *et al.* [1991]. Il est également possible d'avoir recours à des méthodes discriminatives, qui proposent des séparateurs pour lesquels une marge doit être maximisée. Ainsi, dans Harchaoui *et al.* [2009] nous avons proposé de résoudre deux tâches de segmentation ou détection de changements : la segmentation en tours de parole et la segmentation en grandes classes audio. Cette

étude a été réalisée sur deux émissions du talk show *Le Grand Échiquier* et donne des scores de 72% de précision et 63% de rappel pour la segmentation en classes sémantique et 89% de précision et 90% de rappel pour la segmentation en tours de parole. De plus, il est très intéressant de noter que la technique proposée est très compétitive car elle obtient des scores très proches de ceux réalisés avec une méthode supervisée basée sur des modèles de Markov cachés (HMM, voir Rabiner [1989]).

3.4 La détection de concepts de haut-niveau

Dans la figure 3.1, nous avons distingué clairement méthodes non-supervisées et supervisées. Ces dernières comportent deux phases : l'entraînement et le test. La première phase consiste à générer un classifieur sur des données étiquetées, dites d'apprentissage. Ensuite, une fois celui-ci créé, on classe les données de test non-étiquetées, c'est-à-dire pour lesquelles aucune information n'est disponible. La phase d'apprentissage est de fait critique car elle doit permettre une bonne classification des données d'entraînement et de test. Il est par conséquent crucial d'éviter le cas où le classifieur n'est pas assez générique et « sur-apprend » les données. La détection de concept de haut-niveau comme musique, parole, applaudissements, rires, etc. peut être réalisée à l'aide de méthodes supervisées.

Pour la discrimination parole/non-parole, pour laquelle les parties de non-parole réfèrent la plupart du temps à des interruptions musicales et/ou à des bruits environnementaux, Saunders [1996] et Scheirer et Slaney [1997] se basent par exemple sur des modèles de mélanges gaussiens (GMM) en proposant pour chaque classe à discriminer un modèle appris sur les données d'apprentissage. De façon similaire, des classifieurs à machine de vecteurs de support (SVM voir Cristianini et Shawe-Taylor [2000]) peuvent également être utilisés comme cela est proposé par Richard *et al.* [2007]. Dans ce dernier cas, les auteurs utilisent une base de données constituée de flux radiophoniques qui, s'ils ne sont pas à proprement parler exactement identiques au contenu audio d'émissions de talk show, n'en présentent pas moins de fortes similitudes. Leur méthode obtient une F-mesure de 96.5% pour la discrimination parole/musique. De manière similaire, les parties musicales des émissions de talk show peuvent être détectées à l'aide de classifieurs SVM. Les résultats obtenus par le système proposé par Cyril Hory (Télécom ParisTech) dans le cadre du projet francilien Infom@gic confirment l'efficacité de telles méthodes. Ils sont détaillés dans le tableau 3.2 pour quatre émissions du *Grand Échiquier* (on peut se reporter à l'annexe B pour plus d'informations sur les émissions testées).

Il est intéressant de noter que les émissions CPB81052012, CPB81052234 et CPB76068458 comportent beaucoup plus de plages musicales que CPB84052346. Comme en atteste l'annexe B, les deux premières sont en effet consacrées à des musiciens ou orchestre alors que les invités principaux des émissions CPB76068458 et CPB84052346 sont Raymond Devos, un humoriste musicien (piano, tuba, etc. d'où l'importance des interventions musicales) et Gérard Oury, un réalisateur de film.

Certains éléments de structuration audiovisuelle définis comme *délimiteurs* peuvent égale-

émission	durée totale	durée moyenne	nombre	précision	rappel	F-mesure
CPB76068458	78'41"	3'02"	26	97.3%	92.3%	94.7%
CPB81052012	106'38"	2'32"	42	93.5%	90.4%	91.9%
CPB81052234	107'10"	4'07"	26	94.8%	96.8%	95.8%
CPB84052346	43'27"	3'21"	13	99.4%	90.1%	94.5%
moyenne	83'59"	3'15"	27	96.2%	92.4%	94.3%

Tableau 3.2 – Précisions et rappels pour la détection de musique sur quatre émissions du corpus *Le Grand Échiquier* (système Télécom ParisTech implémenté par Cyril Hory). Les durées sont données en minutes/secondes.

ment être détectés par l'analyse du flux audio. Ainsi, il est possible de détecter les rires (voir Wilkins *et al.* [2007]), les jingles (voir Pinquier et André-Obrecht [2004]) ou les applaudissements. Pour ces derniers, Hory et Christmas [2007] proposent une méthode très efficace basée sur l'extraction de nouvelles caractéristiques audio pour la classification de réponses impulsionnelles. Implémentée sur le corpus *Le Grand Échiquier*, celle-ci donne également d'excellents résultats. Le tableau 3.3 donne les précisions et rappels pour quatre émissions :

émission	durée totale	durée moyenne	nombre	précision	rappel	F-mesure
CPB76068458	13'29"	9.8"	83	82.7%	85.5%	84.1%
CPB81052012	9'56"	17.5"	34	89.6%	80.2%	84.6%
CPB81052234	9'55"	21.2"	28	93.9%	88.6%	91.2%
CPB84052346	9'42"	12.7"	46	91.2%	77.6%	83.9%
moyenne	10'45"	15.3"	48	89.4%	83.0%	86.1%

Tableau 3.3 – Précisions et rappels pour la détection d'applaudissements sur quatre émissions du corpus *Le Grand Échiquier* (système Télécom ParisTech implémenté par Cyril Hory). Les durées sont données en minutes/secondes.

L'importance du nombre d'applaudissements pour l'émission CPB76068458 peut s'expliquer par le fait que l'invité principal est l'humoriste Raymond Devos. On peut en effet s'attendre à ce que cette émission comporte beaucoup de sketches et gags provoquant de nombreuses réactions du public et en particulier des applaudissements.

3.5 La détection de concepts de niveau supérieur

Les dernières méthodes que nous présentons dans cet état de l'art concernent la détection de concepts de niveau supérieur. Ceux-ci se définissent comme des événements de niveau « sémantique » plus élevé que les concepts de haut-niveau présentés précédemment. Les événements du type applaudissements, segments musicaux, rires, etc. sont la plupart du temps relativement aisés à détecter puisque montrant une corrélation forte avec le comportement physique d'une des modalités audio ou vidéo. En revanche, pour les concepts de niveau supérieur il est souvent nécessaire d'utiliser des combinaisons de résultats de détections de plus bas niveau.

3.5.1 Quelques exemples

Dans la littérature scientifique de nombreuses études ont été proposées autour de la structuration de retransmissions de compétitions sportives. En particulier, beaucoup de travaux ont été effectués sur les matches de football. Généralement, il s'agit de détecter des actions caractéristiques étiquetées comme « moments forts » (*highlights*) qui correspondent aux moments les plus excitants et intéressants du point de vue du téléspectateur. Celles-ci peuvent par la suite être utilisées pour la création de résumés automatiques. On peut par exemple penser à des actions du type : *but*, *penalty*, *faute*, *carton rouge*, etc. La plupart des travaux existants proposent de détecter de tels événements comme un enchaînement de sous-actions qui peuvt être vu comme un motif caractéristique. Les modèles de Markov cachés sont très fréquemment utilisés pour la détection de ces motifs, que ce soit sur les modalités vidéo (Assfalg *et al.* [2002]), audio (Baillie et Jose [2004]) ou audio et vidéo (Xiong *et al.* [2003]). Certaines études comme Hamid *et al.* [2010] et Kim *et al.* [2010] utilisent le suivi de joueurs et la localisation pour détecter les endroits du terrain ou se produisent les événements d'intérêt. Il est également possible de reconnaître des éléments caractéristiques comme le rond central ou les poteaux de but à l'aide de l'injection de connaissances a priori comme le font Gong *et al.* [1995] et Yu *et al.* [2009] plus récemment.

De façon similaire, d'autres travaux ont proposé des méthodes pour la structucturation de contenus sportifs. Ainsi, pour les matches de tennis, Kijak *et al.* [2006] et Delakis *et al.* [2008] ont exposé des méthodes utilisant des réseaux bayésiens dynamiques (*dynamic bayesian networks* voir Murphy [2002]). Des études ont également été menées sur le basket-ball (Zhou *et al.* [2000] et Xu *et al.* [2003]), le baseball (Zhang et Chang [2002] et Guéziec [2002]), la formule 1 (Petkovic *et al.* [2002]), etc.

La détection de concept de niveau supérieur n'est cependant pas l'apanage des seules émissions sportives. Ainsi dans le cadre de la campagne d'évaluation TRECVid, une tâche d'extraction de descripteurs de haut niveau (*high-level features*) est proposée. Il s'agit de retrouver les occurences d'événements définis à l'avance dans un corpus de vidéos. *Sports*, *Desert*, *US Flag* ou encore *Car* sont des exemples d'éléments à reconnaître. La détection de ceux-ci est rendue difficile par la variation de niveau sémantique. Benmokhtar *et al.* [2007] proposent par exemple d'extraire des attributs audio et vidéo afin d'apprendre un classifieur SVM un contre tous (*one vs all SVM*) pour chaque type d'événement. La décision d'appartenance est ensuite prise en fusionnant les sorties de ces classifieurs à l'aide d'un réseau de neurones.

Pour la détection de publicités certaines études ont été menées. En particulier, Baghdadi *et al.* [2008] ont proposé d'utiliser des réseaux bayésiens dynamiques. Les résultats donnés sont très concluant (90% de précision et 93% de rappel) et laissent penser que l'utilisation de réseaux bayésiens dynamiques peut permettre la détection des composantes principales du talk show.

3.5.2 Études exploratoires pour la détection de concepts audiovisuels

En étudiant le tableau 3.1, on peut remarquer que la détection de certaines composantes structurelles du talk show ne peut pas être réalisée par la seule utilisation des segmenteurs et détecteurs de haut-niveau précédemment présentés. Il faut par conséquent proposer des tech-

niques plus élaborées pour pouvoir appréhender l'information utile contenue dans de tels événements. Des techniques de raisonnement ou d'inférence combinant les sorties de différents classifieurs permettent la détection de concepts tels : les *performances non-musicales*, les *publicités*, les *jingles*, les *inserts*, etc.

Détection d'inserts

Dans le cadre du projet Infom@gic, la tâche de détection d'inserts à été réalisée par Denis Marraud (EADS) en fusionnant les sorties de plusieurs classifieurs. On peut ainsi combiner des détecteurs de bandes noires, caractéristiques de l'essentiel de la production cinématographique lors du passage sur des postes de télévision 4 : 3 ou 16 : 9, des détecteurs de scènes en noir et blanc et des détecteurs de sons environnementaux, typiques des passages pour lesquels les conditions acoustiques générales de la vidéo sont altérées. La détection de ces inserts vidéo est obtenu par fusion des scores de chacun de ces détecteurs (voir Campedel et Hoogstoël [2011] et Bloch [2003]). Comme en atteste le tableau 3.4 celle-ci donne de satisfaisants résultats.

émission	**durée totale**	**durée moyenne**	**nombre**	**précision**	**rappel**	**F-mesure**
CPB76068458	5'49"	2'55"	2	92.2%	45.7%	61.1%
CPB82051110	32'18"	4'02"	8	60.7%	84.4%	70.6%
CPB84052346	54'56"	2'45"	20	84.9%	98.2%	91.1%
moyenne	31'02"	3'14"	10	79.2%	76.1%	77.6%

Tableau 3.4 – Précisions et rappels pour la détection d'inserts sur trois émissions du corpus *Le Grand Échiquier* (système EADS implémenté par Denis Marraud). Les durées sont données en minutes/secondes.

Comme nous l'avons vu précédemment, l'émission CPB84052346 est consacrée au réalisateur Gérard Oury, ce qui explique le nombre très important d'inserts (extraits de films, reportages, etc.). Le faible rappel pour CPB76068458 s'explique au contraire par le nombre très faible d'inserts dans cette émission, deux inserts pour seulement 5 min 49 s.

Détection de performances non-musicales parlantes

Dans une étude préliminaire nous avons exploré la faisabilité de détecter les événements audiovisuels de type *perfomances non-musicales parlantes*, cela sans utiliser de technologies basées sur la transcription automatique de la parole. Ces performances regroupent des interventions du type : lecture d'œuvres, scène de théatre, sketches, etc. mais laissent de côté les performances non-musicales du type numéro de cirque, mimes, etc. ne présentant pas de dialogues. Le défi de cette étude est donc d'être capable de distinguer différents types de registres de langage et en particulier de savoir si la parole traitée fait partie d'une séquence interview ou est une performance, sans avoir le moindre recours à des techniques d'analyse du langage. Comme nous le verrons par la suite, la différence peut parfois être plus tenue qu'on ne le pense. Nous proposons d'élaborer un système basé sur un ensemble de règles heuristiques. Ainsi, une

performance non-musicale parlante contient de la parole, dure au moins trente secondes et est ponctuée d'applaudissements (voir figure 3.3). De plus, le présentateur n'intervient pas lors de celles-ci.

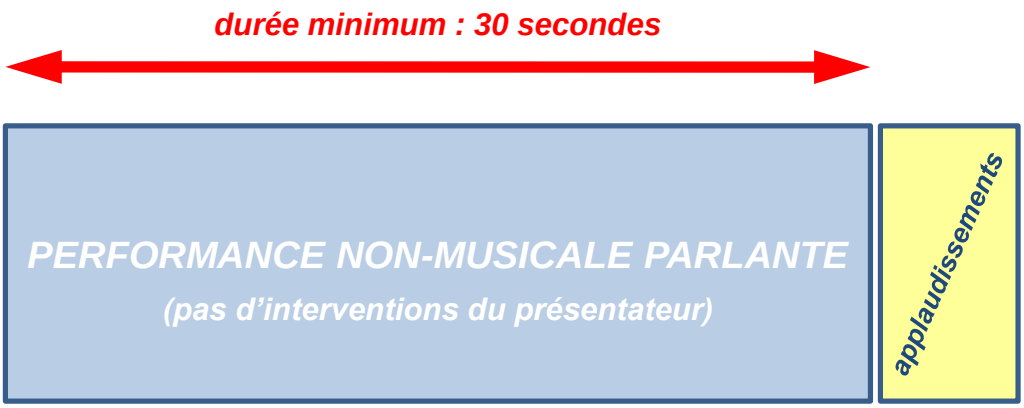


Figure 3.3 – Règles heuristiques caractérisant une performance non-musicale.

Par conséquent, en combinant un détecteur d'applaudissements et des résultats de reconnaissance de locuteurs (dont il sera question dans la partie suivante) on peut vérifier si les hypothèses exposées plus haut sont vérifiées. Les sorties de détecteurs sont moyennées sur des horizons relativement longs en rapport avec la durée moyenne de l'événement. Ainsi, pour la détection d'applaudissements un filtre median d'une longueur de cinq secondes est appliqué sur la sortie du classifieur SVM. Pour la reconnaissance de locuteurs les résultats utilisés sont ceux de Bozonnet *et al.* [2010b] et les sorties sont moyennées sur des fenêtres de deux secondes. Une hypothèse sur le fait que le présentateur, en plus d'être parmis les locuteurs principaux, est la première et/ou la dernière personne à intervenir au cours de l'émission est également utilisée. On obtient alors, sur deux émissions, plusieurs propositions pour des événements de nature *performance non-musicale parlante* (voir figure 3.4). Seules deux émissions ont été traitées ici car aucune des autres émissions traitées par Bozonnet *et al.* [2010b] ne présentent de performances non-musicales parlantes.

À la lueur des segments détectés comme étant des performances non-musicales parlantes, il est intéressant de noter que la limite entre parole sur le plateau d'interview et ce que nous avons défini comme des performances non-musicales est relativement ténue. En effet, les extraits (A) et (C) n'ont par exemple pas été annotés avec le statut de performance non-musicale. Cependant, lorsqu'un invité ou un groupe d'invités prend la parole pendant une trentaine de seconde et est salué à l'issue du discours par les applaudissements du public, cela relève d'un mécanisme de mise en scène qui n'est pas très éloigné de celui observé lors des performances non-musicales parlantes. On pourrait ainsi les qualifier de performances non-musicales parlantes et improvisées.

Figure 3.4 – Performances non-musicales parlantes détectées pour les émissions CPB82055196 (3) et CPB84052346 (2). (A) Michel Sardou parle avec véhémence de l'éviction du PAF (paysage audiovisuel français) de Guy Lux. (B) Mireille Darc lit *Colloque Sentimental* de Verlaine. (C) Pierre Barret parle avec emphase de son amitié avec Michel Sardou. (D) Marie-France Pisier lit une lettre de Colette adressée à la mère du réalisateur Gérard Oury. (E) François Périer et Gérard Oury jouent une scène de *Volpone* de Ben Jonson.

Bilan

Les exemples de détection de concept de niveau supérieur proposés ici indiquent que de fiables descripteurs peuvent être créés en combinant les sorties de classifieurs de plus bas-niveau. L'utilisation de simples règles heuristiques permet déjà l'obtention de bons résultats et il est très probable que l'utilisation de méthodes de type réseaux bayésiens dynamiques puisse encore les améliorer. Cependant, il est important de noter que pour être correctement détectés, les événements audiovisuels proposés sont tributaires des résultats obtenus avec des segmenteurs ou détecteurs de plus bas niveau.

3.6 Vers la reconnaissance de locuteurs

Comme cela a été souligné dans les chapitres 1 et 2, l'information véhiculée par les locuteurs tient une place particulièrement centrale dans les processus de structuration vidéo. De plus, nous avons pu voir que la connaissance de la répartition des tours de parole constituait bien souvent une brique essentielle à la détection de composantes structurelles constitutives du talk show. Par conséquent, contrairement à la vaste majorité des travaux de structuration nous proposons de considérer non pas la segmentation en plans mais les tours de parole comme entités atomiques de structuration. En effet, ceux-ci sont porteurs de beaucoup plus de sens puisque dans le cas des talk shows, l'information utile est véhiculée par les interventions des locuteurs. Il est donc essentiel d'obtenir des informations fiables sur les prises de parole des intervenants, la détection de nombreux composants structurels du talk show reposant sur la compréhension de leur répartition.

Traditionnellement, le terme de *reconnaissance de locuteurs* désigne de manière générique toutes les applications où l'on cherche à obtenir des renseignements concernant l'identité d'une personne à partir d'enregistrements audio. Plusieurs types d'applications entrent dans le cadre de la reconnaissance de locuteurs. On peut ainsi distinguer la *vérification du locuteur* qui permet de décider si l'identité revendiquée par un locuteur est effectivement compatible avec sa voix, de l'*identification du locuteur* pour laquelle le but est de déterminer, parmi un ensemble de locuteurs potentiels, à quel locuteur correspond un enregistrement vocal, etc. Pour ces deux exemples, la reconnaissance de locuteurs repose sur l'apprentissage préalable de caractéristiques vocales des locuteurs.

Dans notre travail, nous avons fait le choix de ne pas considérer cette étape préalable, et par là même, de nous concentrer sur des méthodes complètement non-supervisées. En effet, il est souvent impossible de construire des bases d'apprentissage des caractéristiques des locuteurs potentiellement présents sur un plateau de talk show. Là encore, afin de garantir la généricité de notre schéma de structuration, nous souhaitons limiter au maximum l'injection de connaissance a priori. Par la suite c'est à ce type de systèmes que nous ferons référence en parlant de reconnaissance de locuteurs (*speaker diarization*) même si le terme regroupement de locuteurs pourrait sembler plus approprié. La définition généralement admise dans ce cas est qu'il s'agit d'un processus visant à partitionner un flux audio en segments homogènes en accord

avec l'identité des locuteurs (voir figure 3.5). Cela permet ainsi, dans les limites de performance des méthodes considérées, d'étiquetter d'un même *label* les interventions de chaque individu et de répondre à la question : « Qui parle quand ? », la correspondance entre modèles et identités des locuteurs pouvant être établie ultérieurement par l'utilisateur.

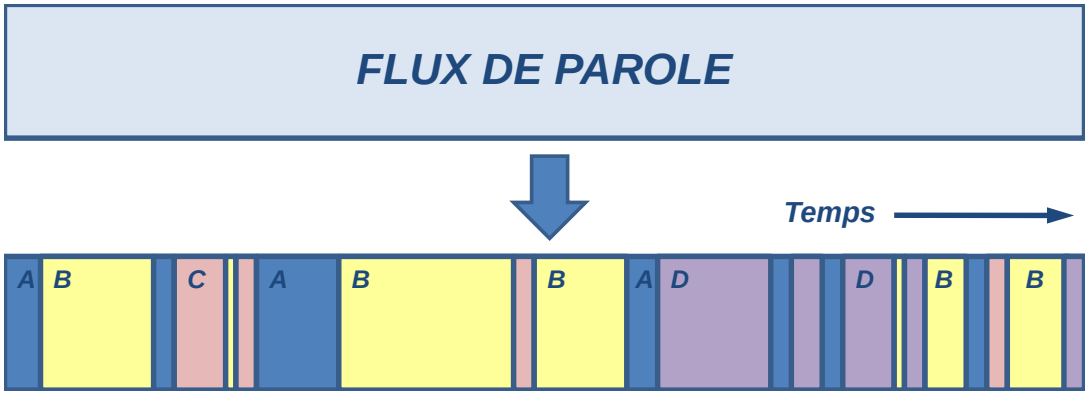


Figure 3.5 – Schéma de principe de la reconnaissance de locuteurs. Il s'agit du découpage du flux de parole en interventions de différents locuteurs (ici *A*, *B*, *C* et *D*)

.

Cet aspect étant d'une importance cruciale dans le travail présenté, nous y consacrons la seconde partie de cette thèse. Ainsi, dans le chapitre 4, nous proposons un tour d'horizon détaillé des techniques existantes. Nous évaluons ensuite les résultats de reconnaissance obtenus avec un système état de l'art lorsqu'appliqué au type de contenu particulier qu'est le talk show.

Conclusion

Suite à la proposition de structuration donnée dans le chapitre 2, nous nous sommes attachés à montrer que les composantes génériques constitutives de tout talk show peuvent être identifiées de façon relativement satisfaisante au moyen de diverses techniques. Ainsi nous avons présenté différentes méthodes issues de l'état de l'art ainsi que les résultats de quelques études préliminaires pointant dans cette direction. Une distinction importante a été effectuée entre les méthodes de segmentation, de détection de concept de haut-niveau et de niveau supérieur. Les résultats fournis dans ce chapitre ont majoritairement été obtenus dans le cadre du projet de recherche Infom@gic sur des sous-ensembles du corpus *Le Grand Échiquier*.

Ainsi, nous avons observé que des résultats satisfaisants pouvaient être obtenus pour des composantes structurelles relativement proches du signal (applaudissements, musique, rires, etc.) ainsi que pour des concepts porteurs de plus d'information utile comme les inserts ou les performances non-musicales parlantes. La détection de ces éléments a mis en avant la néces-

sité d'avoir à disposition une identification de qualité des interventions des différents locuteurs présents sur le plateau. Par conséquent nous nous concentrons sur cet aspect dans la seconde partie, celui-ci étant essentiel dans le travail que nous décrivons.

NOUS avons vu au cours de cette partie que les émissions de talk show, malgré leurs différences, présentent un nombre important de caractéristiques communes. En particulier, des études sémiologiques ont montré l'importance du couple présentateur(s)/invité(s) sur lequel repose ces programmes. À la lueur de cas d'usage, une structure générique de ce type de contenu a également pu être proposée. La validité de celle-ci a ensuite été confirmée par une expérience utilisateur soulignant les bénéfices d'une telle organisation pour une tâche particulière : la recherche d'extraits audiovisuels.

Ayant adopté cette organisation, nous avons proposé un aperçu des techniques à mettre en œuvre pour détecter automatiquement les composantes génériques constitutives de celui-ci. Appliquées au corpus télévisuel *Le Grand Échiquier* nous avons détaillé des approches de segmentation possibles pour les éléments de « bas niveau ». Nous avons également présenté quelques résultats pour la détection de concepts de haut-niveau et de niveau supérieur, c'est à dire pour des événements véhiculant des informations de niveau « sémantique » plus élevé. Par exemple, des détecteurs d'inserts et de performances non-musicales parlantes ont été proposés.

Enfin, le rôle de la reconnaissance de locuteurs a été soulevé. Celui-ci est particulièrement crucial dans l'organisation automatique que nous proposons puisque nous basons le tour de parole comme entité fondamentale de structuration. Par conséquent, nous proposons dans la seconde partie de cette thèse un panorama des méthodes proposées par la communauté scientifique pour la résolution de cette tâche ainsi qu'une architecture originale de reconnaissance multimodale.

Deuxième partie

Reconnaissance multimodale de locuteurs

APRÈS avoir exposé les enjeux de la structuration audiovisuelle et en particulier la structuration d'émissions de talk show dans la partie précédente, nous nous sommes focalisés sur la proposition d'une organisation interne de ce type d'émissions. L'identification des interventions des différents locuteurs présents sur le plateau s'est alors avérée cruciale. En effet, dans notre schéma de structuration, le tour de locuteur joue le rôle de brique élémentaire ou entité atomique.

De fait, après avoir exposé diverses techniques permettant l'identification des éléments de structuration proposés, nous avons souligné l'importance d'obtenir des résultats de reconnaissance de locuteurs suffisamment fiables. Par conséquent, nous proposons ici un tour d'horizon détaillé des différentes méthodes de reconnaissance de locuteurs antérieures avant de presenter diverses améliorations exploitant la modalité image, puis, une architecture originale de reconnaissance multi-modale. Celle-ci constitue une des contributions principales de cette thèse.

Chapitre 4

État actuel des méthodes pour la reconnaissance de locuteurs

Introduction

Nous avons vu dans la première partie que les locuteurs occupent une place de choix parmi les éléments dont la détection est nécessaire pour s'assurer d'une bonne organisation automatique de talk show. En effet, ces derniers portent la plus grande partie de l'information « humaine » ou « sémantique ». De plus, nous avons largement insisté sur le fait que nous souhaitons nous affranchir de système de transcription de la parole afin de garantir la généricité de notre approche de structuration. C'est précisément pour cette raison que le parti pris de ce travail de thèse a été de placer les locuteurs, et plus précisément les tours de parole, au centre du processus d'organisation du contenu alors que la majorité des travaux de structuration audiovisuelle utilisent le plan (*shot*) comme unité atomique. Par conséquent, il est important d'effectuer un tour d'horizon des méthodes de reconnaissance de locuteurs proposées dans l'état de l'art et d'évaluer leurs apports respectifs dans le cadre d'un travail de structuration de talk show.

4.1 Qu'est ce que la reconnaissance de locuteurs ?

La reconnaissance de locuteurs est composée de plusieurs étapes : détection de parole, détection de changement de locuteurs et regroupement (ou *clustering*) de locuteurs (voir figure 4.1) qui seront détaillées par la suite. Celles-ci, bien qu'étant isolément des domaines de recherche à part entière, peuvent être considérées comme des blocs élémentaires aux méthodes de traitement de la parole, qu'il s'agisse d'indexation de locuteur, de reconnaissance de locuteurs, de transcription automatique de parole, de traduction automatique, etc.

De plus il est bon de noter que la reconnaissance de locuteurs est un domaine qui a bénéficié largement de la mise en place de campagnes d'évaluation à grande échelle permettant aux différentes équipes de recherche de mesurer les performances de leurs systèmes sur des bases de données communes. Ainsi, depuis les années 1980, le *National Institute of Standards and Technology*[1] (NIST) propose des évaluations visant à promouvoir et comparer les avancées en

1. agence du Département du Commerce des États-Unis favorisant l'économie des technologies et des standards

matières de traitement de la parole. En particulier depuis 2002, les évaluations *Rich Transcription* (RT) mettent l'accent sur un certain nombre de tâches liées à ce domaine. À l'origine effectuées sur des données téléphoniques, ces évaluations ont également eu pour objet les bulletins d'informations (*broadcast news*) et les réunions de travail (*conference meetings*). Parallèlement, d'autres campagnes d'évaluations ont été menées comme l'*évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophonique* (ESTER, voir Galliano *et al.* [2009]) mise en place par l'Association Française de la Communication Parlée (AFCP).

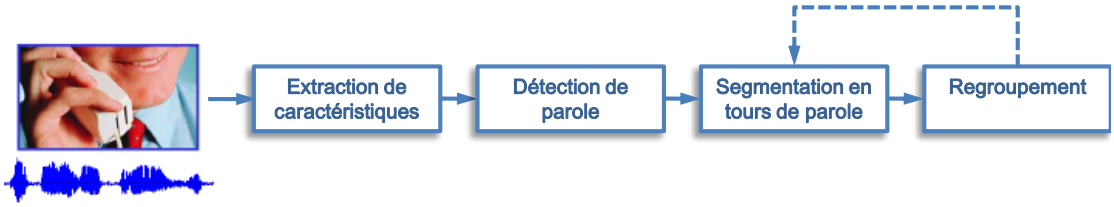


Figure 4.1 – Schéma illustrant les étapes constitutives de la reconnaissance de locuteurs (*speaker diarization*). Le trait pointillé indique que certaines méthodes proposent d'effectuer conjointement et itérativement segmentation et regroupement.

Historiquement, les premières méthodes proposées utilisaient exclusivement la modalité audio. Cependant, depuis quelques années, les corpus à traiter ont évolué faisant la part belle aux données vidéo. De nouvelles méthodes basées sur l'analyse conjointe des deux modalités ont donc vu le jour proposant ainsi des travaux inspirés des approches biométriques.

4.2 Les étapes de la reconnaissance de locuteurs

Bien que des approches multimodales aient été proposées depuis (comme présenté dans la section 4.3), les techniques de reconnaissance de locuteurs se basent sur l'utilisation quasi-exclusive de données audio. En général, les données acoustiques utilisées sont les MFCC (*Mel-Frequency Cepstral Coefficients*), des coefficients cepstraux calculés par l'application d'une transformée en cosinus discrète au spectre de puissance d'un signal. Les bandes de fréquence de ce spectre sont espacées logarithmiquement selon l'échelle perceptive de Mel (voir Rabiner et Juang [1993]). Cependant, on trouve parfois d'autres types de descripteurs comme les coefficients linéaires perceptifs (PLP) ou des caractéristiques prosodiques (comme proposé par Friedland *et al.* [2009b]).

4.2.1 Approches agglomératives et divisives

La grande majorité des travaux de reconnaissance de locuteurs peut être divisée en deux catégories : les approches agglomératives (*bottom-up*) et divisives (*top-down*) comme cela est proposé par Anguera [2006], Tranter et Reynolds [2006] et Anguera *et al.* [2011]. Les premières

sont initialisées avec un nombre important de *clusters*, représentant des modèles de parole. Ce nombre est généralement plus grand que le décompte réel des locuteurs présents. Les secondes, au contraire, partent avec moins de *clusters* (la plupart du temps un) qui sont ensuite subdivisés à chaque itération (voir figure 4.2). Dans les deux cas, il s'agit de converger vers le nombre optimal de *clusters* à savoir le nombre réel de locuteurs. Si ce nombre n'est pas atteint, on parle de *sous-clustering* (moins de *clusters* que de locuteurs réels) ou de *sur-clustering* (plus de *clusters* que de locuteurs réels). La plupart de ces méthodes sont basées sur des modèles de Markov à états cachés (HMM, voir Rabiner [1989]) où chaque état correspond à un modèle de locuteur et est représenté par un mélange de gaussiennes (GMM, voir Duda *et al.* [2000]) et où les transitions entre les états représentent les passages d'un locuteur à un autre (ou tours de parole).

Les méthodes agglomératives (*bottom-up*) sont les plus représentées dans l'état de l'art de la reconnaissance de locuteurs. Elles consistent en l'agrégation d'un nombre important de modèles par le biais d'un regroupement ou *clustering* hiérarchique agglomératif comme indiqué figure 4.3. Ce regroupement se produit généralement itérativement et de nouveaux modèles sont appris à la suite de chaque fusion de *clusters*. Plusieurs méthodes ont été proposées pour l'initialisation. En général, le flux audio est sur-segmenté en un nombre important de *clusters*, que ce soit de façon uniforme avec un nombre fixé de segments de longueur équivalente, à l'aide de techniques de *pré-clustering* ou par identification des tours de parole. Le regroupement de *clusters* s'effectue ensuite jusqu'à ce qu'un critère d'arrêt soit atteint. De nombreuses méthodes ont été proposées comme le critère d'information bayesienne (BIC, Reynolds et Torres-Carrasquillo [2005] ou Wooters et Huijbregts [2008]), la distance de Kullback-Leibler (KL, Rougui *et al.* [2006]) ou encore le rapport de vraisemblance généralisée (GLR, Tsai *et al.* [2004]).

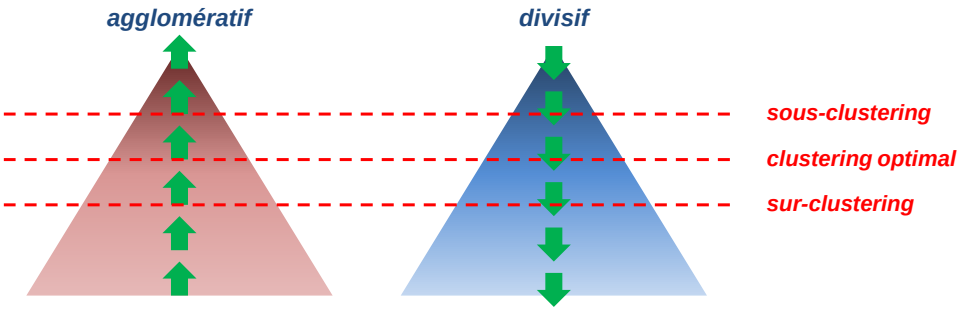


Figure 4.2 – Schéma général pour les architectures agglomératives (*bottom-up*) et divisives (*top-down*).

De leur côté, les méthodes divisives (*top-down*) proposent de commencer par modéliser les parties de parole avec un seul modèle de locuteur puis d'ajouter continuellement de nouveaux modèles jusqu'à ce que le nombre de locuteurs attendu ou un critère d'arrêt soit atteint. Une fois qu'un mélange de gaussiennes est appris sur tous les segments de parole disponibles, une procédure de sélection identifie des données permettant l'apprentissage itératif de nouveaux modèles de locuteurs qui sont ensuite rajoutés au fur et à mesure au modèle initial. Chaque subdivision est entrecoupée d'un décodage de Viterbi et d'une adaptation par maximum a posterori (MAP)

ou d'un entraînement espérance maximisation (EM). Les segments attribués aux nouveaux modèles passent alors du statut de non-étiquetés à celui d'étiquetés et ne peuvent plus être utilisés pour la création d'un nouveau modèle à l'itération suivante. L'algorithme s'arrête ensuite, soit naturellement car il ne reste plus de segments non-étiquetés, soit lorsqu'un critère d'arrêt est atteint. Les critères sont bien souvent les mêmes que ceux utilisés avec les méthodes *bottom-up*. De tels systèmes peuvent être trouvés dans les travaux de Meignier *et al.* [2001] ou encore de Fredouille *et al.* [2009]. Généralement un peu moins performants que les meilleurs systèmes *bottom-up*, ceux-ci présentent les avantages d'une implémentation plus légère.

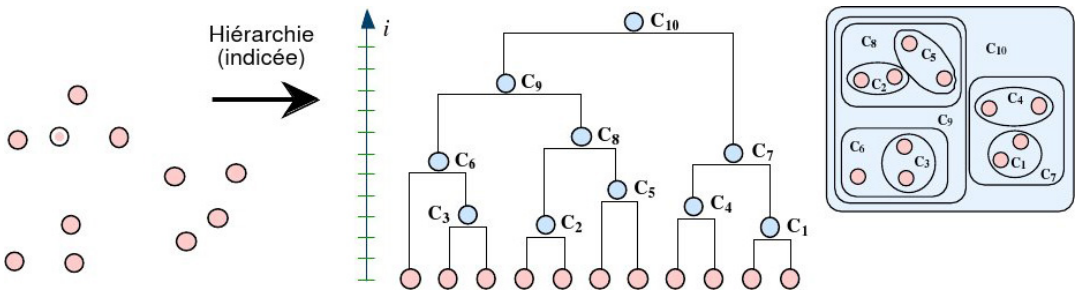


Figure 4.3 – Classification par regroupement hiérarchique. Les *clusters* les plus proches en terme de distance sont appariés. L'arbre figurant au milieu rend compte des différents appariements et est appelé dendrogramme.

Après avoir détaillé les deux philosophies principales des systèmes de reconnaisance de locuteurs, nous pouvons nous intéresser de plus près à leurs étapes constitutives indiquées sur la figure 4.1 ainsi qu'aux différentes méthodes qui ont été proposées dans la littérature.

4.2.2 La détection automatique de la parole

La reconnaissance de locuteurs suppose une capacité à bien discriminer les zones de parole pour pouvoir ensuite les traiter efficacement. On parle alors de détection automatique de la parole (*speech activity detection*). Il s'agit d'écarter tous les segments audio de *non-parole* afin d'éviter la création de modèles acoustiques impurs, ce qui aurait des conséquences fâcheuses lors des étapes ultérieures. Initialement, les premiers sytèmes tentaient d'écarter les données de non-parole en modélisant un *cluster* pour celles-ci lors du processus d'agrégation/subdivision. Cependant, il a été rapidement observé que de meilleurs résultats pouvaient être obtenus par l'ajout de l'étape de classification parole/non-parole (*speech/non-speech*).

De nombreuses approches ont été présentées pour effectuer la discrimination parole/non-parole (voir Ramírez *et al.* [2007]). Cette tâche est relativement ancienne et des recherches ont été effectuées dans d'autres circonstances que la seule reconnaissance de locuteurs, en particulier pour la discrimination parole/musique (voir Saunders [1996], Scheirer et Slaney [1997] et Williams et Ellis [1999]). La classification parole/non-parole est en général réalisée par classification par maximum de vraisemblance sur des modèles de mélanges de gaussiennes (GMM)

qui sont préalablement entraînés sur des données d'apprentissage. D'autres modèles comme les modèles de Markov à états cachés peuvent également y être ajoutés (Wooters *et al.* [2004] et Wooters et Huijbregts [2008]). Certains systèmes, comme celui de Richard *et al.* [2007] réalisé dans le cadre de la campagne ESTER, proposent d'utiliser des classifieurs SVM (machines à vecteurs de support) pour discriminer les différentes classes audio (dans ce cas sur des données de type radiophoniques). L'avantage de créer des modèles autres que celui de parole est de minimiser le taux de réjection de véritable segments de parole, ce qui peut arriver lorsque parole et bruit ou parole et musique sont mélangés.

4.2.3 La segmentation en tours de parole

Cette étape a pour but d'identifier les points de rupture du flux audio qui témoignent d'un changement de locuteur. Comme nous l'avons vu auparavant, ces points peuvent être qualifiés de tours de parole et la segmentation en locuteurs résulte de leur détection. Dans le cas où le flux à analyser n'est pas constitué de parole (ou pas uniquement) on parle de détection de changements ou de rupture (voir par exemple Désobry *et al.* [2005]). La quasi-totalité des approches de détection de changements procède de manière similaire. Deux fenêtres d'observation adjacentes se déplacent à un pas d'avancement donné et des mesures de distance entre les deux fenêtres sont calculées.

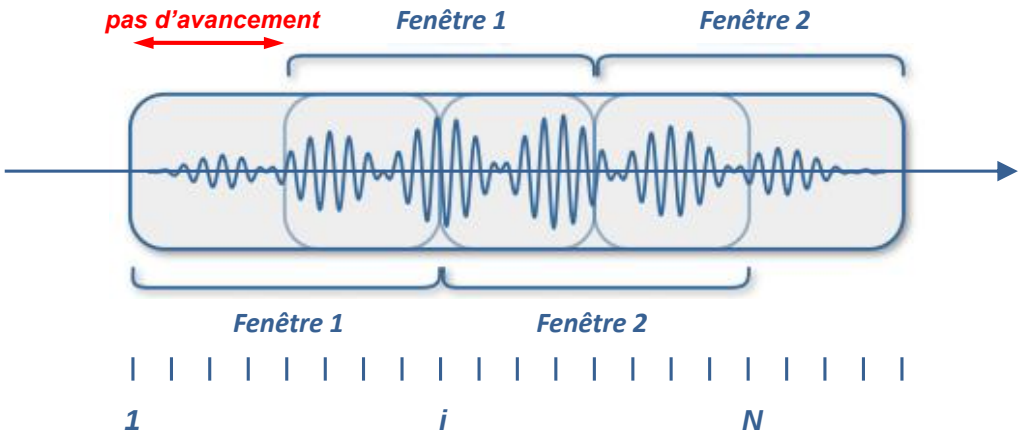


Figure 4.4 – Fenêtres d'analyses glissantes avec un pas d'avancement de la moitié de la taille d'une fenêtre.

Cela permet de décider à chaque pas si un changement physique, reflété par un changement dans la distribution des données, est détecté. La figure 4.4 donne un aperçu de ces fenêtres glissantes. Une des méthodes les plus répandues pour cette tâche est l'étude des variations du critère d'information bayesienne (BIC) initialement introduit par Chen et Gopalakrishnan [1998].

Celui-ci consiste à modéliser les données de chaque fenêtre par des gaussiennes multiva-

riées et à comparer si celles-ci sont mieux modélisées par une seule (pas de changement) ou deux distributions (changement) :

$$X_i \sim N(\mu_i, \Sigma_i)$$

Deux hypothèses sont alors envisageables (en considérant $N(\mu, \Sigma)$ une loi normale de moyenne μ et variance Σ) :

$$\mathbf{H_0} : X_1 ... X_N \sim N(\mu, \Sigma) \qquad \text{et} \qquad \mathbf{H_1} : X_1 ... X_i \sim N(\mu_1, \Sigma_1) \text{ et } X_{i+1} ... X_N \sim N(\mu_2, \Sigma_2)$$

Le critère BIC à un instant i est le suivant :

$$BIC(i) = R(i) - \lambda P \tag{4.1}$$

avec $R(i) = N \log|\Sigma| - N_1 \log|\Sigma_1| - N_2 \log|\Sigma_2|$, le rapport de vraisemblance

et $P = \frac{1}{2}(d + \frac{1}{2} d(d+1)) \log N$, une pénalisation,

d étant la dimension de l'espace des descripteurs audio, λ un poids à fixer et N_k le nombre d'observations dans la fenêtre k ($N_1 + N_2 = N$). Ainsi si le critère BIC est positif, c'est l'hypothèse $\mathbf{H_1}$, indiquant un changement, qui est favorisée, sinon, c'est $\mathbf{H_0}$. Le problème des méthodes de segmentation de type BIC est qu'elles nécessitent de spécifier le seuil λ. Elles sont néanmoins utilisées dans un grand nombre de travaux et certaines méthodes proposent une adaptation automatique du paramètre λ (voir Ben *et al.* [2004] ou Anguera *et al.* [2006] par exemple).

Il existe des alternatives aux méthodes BIC comme le rapport de vraisemblance généralisé (GLR pour *gaussian likelihood ratio*) présenté par Gish *et al.* [1991], Delacourt et Wellekens [2000] et Khoury *et al.* [2008] ou encore la divergence de Kullback-Leibler (KL), même si on préfère souvent à cette dernière sa version symétrique KL2 (voir Siegler *et al.* [1997] et Barras *et al.* [2006]).

Enfin, bien que généralement peu employées, il existe encore d'autres méthodes que celles présentées plus haut, en particulier celles basées sur l'utilisation de méthodes à noyaux. En raison de leur coût computationnel, leur rayonnement dans la communauté de traitement de la parole reste plus confidentiel. On peut par exemple citer les travaux de Fergani *et al.* [2008] qui comparent les distributions des deux fenêtres glissantes au moyen de machines à vecteur de support à une classe (*one-class SVM*). De façon similaire, Harchaoui *et al.* [2009] proposent un critère de détection de nouveauté basé sur le discriminant de Fisher dans l'espace kernelisé (KFDA).

4.2.4 Le regroupement de locuteurs

L'étape de regroupement de locuteurs ou *clustering* a pour but l'association des segments de parole identifiés par la segmentation avec le locuteur correspondant. Idéalement, le résultat final doit aboutir à la création d'un *cluster* par locuteur. La dissimilarité entre les segments est donc mesurée à l'aide de métriques type BIC (Moraru *et al.* [2005]), KL (Rougui *et al.* [2006]), KL2 (Ben *et al.* [2004]), CLR (*cross likelihood ratio*, Khoury *et al.* [2008]), etc.

Cependant, le problème rencontré avec les méthodes précédentes est l'impossibilité de couper des segments et de réassigner les trames qu'ils contiennent (et par là-même la dépendance à d'éventuelles erreurs d'initialisation). Par conséquent, la majorité des travaux de reconnaissance de locuteurs effectuent de manière conjointe segmentation et regroupement. C'est ce que proposent Ajmera et Wooters [2003], Wooters et Huijbregts [2008] et Friedland *et al.* [2009b] pour les algorithmes de type *bottom-up* et Meignier *et al.* [2001] et Fredouille *et al.* [2009] pour ceux de type *top-down*. Dans chaque cas, des modèles GMM/HMM sont utilisés. Après un entraînement des modèles par des méthodes de type espérance maximisation (EM, voir Fredouille *et al.* [2009]) ou maximum a posteriori (MAP, voir Ben *et al.* [2004]), un décodage de Viterbi est réalisé pour effectuer une reclassification des données de parole. Puis un nouveau regroupement peut avoir lieu, la boucle se répétant jusqu'à stabilisation de la sortie reclassée.

Enfin, d'autres mesures ont encore été proposées pour l'agrégation hiérarchique des *clusters*. Dans leurs travaux, Fergani *et al.* [2008] proposent de calculer une mesure de dissimilarité dans l'espace kernelisé (*kernel-based dissimilarity measure*).

4.3 Les approches multimodales

Relativement récemment, plusieurs méthodes ont été proposées alliant l'utilisation de données audio et vidéo pour la reconnaissance de locuteurs (Friedland *et al.* [2009a] par exemple). Ces travaux n'ont pu voir le jour que grâce à la création de bases de données audiovisuelles de grandes tailles suppléant celles uniquement audio des campagnes NIST et ESTER. Ainsi, le corpus AMI, libre de droits, met à la disposition des chercheurs des réunions de travail filmées sous plusieurs angles, enregistrées avec des réseaux de microphones et au cours desquelles les intervenants suivent différents scénarii (voir Carletta *et al.* [2005]). De façon identique, Canal9 est une base de données de débats politiques filmés d'environ 43 heures (72 émissions). Plus de détails sont donnés par Vinciarelli *et al.* [2009]. Ces corpus sont présentés par la figure 4.5.

Ces deux bases se distinguent par leur contenu vidéo : brut dans le cas du corpus AMI (c'est-à-dire avec les images simultanées de chaque caméra) et monté dans celui de Canal9 (résultat de choix d'un réalisateur). À notre connaissance, il n'existe pas de travaux de reconnaissance multimodale de locuteurs antérieurs à la création de ces deux bases. Cependant, des travaux d'intérêt croisant modalités audio et vidéo ont déjà été menés dans le champ d'études biométriques axées sur les techniques de vérification de l'identité.

Figure 4.5 – Captures d'écran des corpus de réunions de travail AMI (en haut à gauche) et de débats politiques Canal9 (en bas à droite).

4.3.1 À l'origine, les travaux biométriques

S'il a fallu attendre relativement longtemps avant que ne soient disponibles des bases de données multimodales pour la reconnaissance de locuteurs, des travaux exploitant simultanément les modalités audio et vidéo ont été proposés beaucoup plus tôt. En effet, de nombreuses bases de données biométriques ont permis l'essort de ces méthodes. On peut par exemple citer BANCA (Biometric Access Control for Networked and e-Commerce Applications), CUAVE (Clemson University Audio-Visual Experiments, voir Patterson *et al.* [2002]), MBGC (Multiple Biometric Grand Challenge), MVGL-AVD (Multimedia, Vision and Graphics Laboratory Audio-Visual Database) ou encore XM2VTSDB (The Extended Multi Modal Verification for Teleservices and Security applications DataBase). Ces corpus ont été créés pour des utilisations exclusivement biométriques, à savoir que les visages sont orientés de face, que la parole est généralement intelligible (sauf si elle est artificiellement bruitée), qu'il y a un nombre limité de participants par vidéo (au maximum deux ou trois), etc. Dans un scénario réel, le discours des locuteurs serait plus libre, les mouvements des sujets ne pourraient pas être contôlés et par conséquent certaines détections, comme celles du visage ou de la bouche, pourraient poser problème.

Nock *et al.* [2003] et Bredin et Chollet [2006] présentent un état de l'art conséquent sur les différentes méthodes qui ont été proposées pour répondre aux questions d'identité du locuteur. Certaines méthodes comme celle exposée par Hershey et Movellan [2000] utilisent l'information mutuelle ou le calcul d'un coefficient de corrélation de Pearson pour identifier les pixels de la vidéo les plus à mêmes de correspondre au signal audio et localiser ainsi un locuteur actif. Slaney et Covell [2000] et Sargin *et al.* [2007] proposent eux de procéder à une analyse de corrélation canonique (CCA), permettant de projeter les données audio et vidéo sur une base jointe

maximisant leur corrélation et fournissant ainsi une mesure de synchronie. Des méthodes non-paramétriques sont proposées par Fisher *et al.* [2000] et Fisher et Darrell [2001] pour identifier les mouvements des lèvres et vérifier ainsi si une personne est effectivement un locuteur actif alors que Bredin et Chollet [2007] utilisent une méthode hybride CCA/CoIA pour réaliser la même tâche. L'analyse de co-inertie, CoIA, étant une transformation similaire à CCA où, à la place de la corrélation, la covariance est maximisée. Enfin, l'analyse factorielle cross-modale (CFA) est étudié dans Li *et al.* [2003] et présente pour la détection de visages parlants des avantages de résistance au bruit par rapport à l'analyse de corrélation canonique.

4.3.2 La reconnaissance multimodale de locuteurs

La constitution des bases de données AMI et Canal9 (voir figure 4.5) a permis l'émergence de nouveaux travaux multimodaux basés cette fois sur des données « réelles ». Pour le corpus Canal9, Dielmann [2010] propose de détecter les scènes de même nature, à savoir même locuteur et même disposition sur le plateau, et de les regrouper pour présenter à un potentiel utilisateur une vue ramassée de l'émission traitée. Parallèlement, les études menées sur des vidéos de reunions de travail (*meeting conference*) par Vajaria *et al.* [2006] et Hung et Friedland [2008] mettent en évidence le bien-fondé de s'appuyer sur des descripteurs plus robustes que le simple mouvement des lèvres. Ils proposent par exemple de s'intéresser aux mouvements de la tête ainsi qu'aux mouvements globaux du corps des intervenants, ayant montré que généralement lorsque ceux-ci parlent, ces mouvements sont également très corrélés avec le flux audio. Outre les vidéos de réunions de travail et les débats télévisés, les émissions de talk show ont également été l'objet de plusieurs travaux d'intérêt. Dans un contexte plus proche de celui de cette thèse, Bendris *et al.* [2010b] et Bendris *et al.* [2010a] effectuent des mesures d'entropie sur l'activité labiale des personnes à l'écran pour l'émission de talk show *On n'a pas tout dit*, leur permettant ainsi d'effectuer une indexation des participants.

C'est cependant seulement depuis les travaux de Friedland *et al.* [2009a] et Friedland *et al.* [2009c] qu'on peut parler de façon effective de systèmes de reconnaissance audiovisuelle de locuteurs. Basée sur le corpus AMI, la première approche repose sur l'utilisation de plusieurs prises de vue (mode multi-caméra) alors que dans la seconde, une seule est utilisée. Le processus de reconnaissance de locuteurs est effectué en utilisant un système similaire à celui de Wooters et Huijbregts [2008] mais en créant cette fois pour chaque *cluster* deux modèles GMM (un pour chaque modalité). De plus, il s'est avéré que l'utilisation conjointe des données audio et vidéo permet de déterminer la position dans l'image des locuteurs actifs et ainsi de répondre à la question « Qui parle quand et où ? ».

4.4 Comparaison entre émissions de talk show et enregistrements de réunions de travail

Les systèmes de reconnaissance de locuteurs sont fréquemment développés pour des données de réunion de travail, l'objet d'étude depuis plusieurs années des campagnes d'évaluation

NIST RT. Les bases de données de réunions de travail sont généralement enregistrées au moyen de microphones isolés ou en réseaux, posés sur une table centrale. Le rapport signal sur bruit est donc en général plus élevé que lorsque les intervenants sont munis de micro-cravates puisque la configuration du lieu d'enregistrement, sa réverbération ainsi que la distance aux microphones interviennent grandement. En effet, le bruit de fond capté dans ces cas-là est alors plus important, la qualité des enregistrements peut être dégradée, les niveaux de volume peuvent varier, etc. De plus, le type de discours produit lors des réunions de travail des bases de données NIST RT est en général très spontané avec de nombreux passages de double-voix, c'est-à-dire pendant lesquels deux intervenants s'expriment en même temps.

Les émissions de talk show présentent des caractéristiques communes avec les bases de données de réunions de travail. En effet, là aussi les conditions acoustiques sont difficiles. Par exemple, pour *Le Grand Échiquier*, on observe que les microphones utilisés sont à la fois des micro-cravates et des micros fixes (voir figure 4.6). De plus, en raison de la faible qualité d'enregistrement (voir annexe A), le signal audio est assez fortement dégradé. Enfin, comme nous l'avons explicité dans notre présentation des programmes de talk show du chapitre 1, le discours des intervenants est ici aussi très spontané et riche en phénomènes de double-voix.

Figure 4.6 – Positions des micros fixes (en rouge) et cravates (en vert) sur le plateau de l'émission CPB82055196 du corpus *Le Grand Échiquier*.

Cependant, en plus de la présence d'inserts, de performances musicales, d'applaudissements ou de rires, les silences observés généralement à l'occasion des tours de parole sont extrêmement faibles voir négligeables. En comparaison avec les réunions de travail au cours desquelles les intervenants font fréquemment des pauses afin d'éclaircir leurs idées ou de prendre le temps de répondre à une question, les programmes de talk show présentent une parole plus vive et parfois quasiment écrite. Cela est surement dû au fait que les thèmes développés sont généralement préparés à l'avance par le présentateur et ses invités.

Le tableau 4.1 propose quelques éléments de comparaison entre réunions de travail et émissions de talk show. Il est intéressant de noter que les émissions du corpus *Le Grand Échiquier* sont beaucoup plus longues — en moyenne 2 h 27 min contre 25 min — que celles du corpus NIST RT'09. Une fois les segments de bruit et musique enlevés restent alors 50 minutes de parole à traiter contre 13. Concernant la parole cependant, il semble que celle-ci soit fortement spon-

tanée avec dans les deux cas de nombreux phénomènes de double-voix. De plus, on observe beaucoup plus de locuteurs pour *Le Grand Échiquier* que pour les émissions de réunions de travail, ce qui peut être attendu étant donnée la différence de longueur. En revanche, on peut noter une très grande disparité au niveau des temps de parole. Il s'agit véritablement là d'un challenge supplémentaire concernant le traitement des émissions de talk show. En effet, les locuteurs majoritaires ont tendance à écraser le reste des intervenants en rendant l'apprentissage de modèles pour les seconds très délicat.

événement	***Le Grand Échiquier***	**NIST RT'09**
nombre d'émissions	7	7
durée d'une émission	147'	25'
temps de parole	50'	13'
nombre de segments	1033	882
longueur moyenne d'un segment de parole	3"	2"
temps de double-voix	5'	3'
nombre de locuteurs	13	5
temps de parole du locuteur le plus actif	24'36"	8'55"
temps de parole du locuteur le moins actif	7"	2'26"

Tableau 4.1 – Statistiques de parole pour 6 émissions du *Grand Échiquier* et les 7 émissions de la base de données de réunions de travail NIST RT 2009 (d'après Bozonnet *et al.* [2010b]). Les durées sont données en minutes/secondes.

4.5 Les méthodes d'évaluation

Dans le cadre des campagnes NIST et ESTER des métriques particulières ont été créées pour pouvoir comparer les performances des différents algorithmes soumis. Les résultats de reconnaissance de locuteurs fournis par les participants à ces campagnes doivent inclure les temps de début et de fin pour chaque segment de parole, accompagnés de l'étiquette correspondant au locuteur actif identifié. En comparant avec la segmentation cible (ou de référence), un taux d'erreur de reconnaissance (TER) peut alors être calculé (*diarization error rate*).

La métrique la plus couramment utilisée est celle proposée par NIST [2] qui mesure la fraction de temps qui n'est pas correctement attribuée à un locuteur ou qui n'est pas de la parole. Puisque la tâche ne se définit pas par l'identification des locuteurs, la première étape est d'effectuer un alignement des *clusters* proposés avec les locuteurs véritables de la segmentation de référence. Le taux d'erreur de reconnaissance se calcule comme suit :

$$TER = \frac{\sum_{s=1}^{S} dur(s) \cdot \left(max\left(N_{ref}(s), N_{sys}(s)\right) - N_{correct}(s)\right)}{T_{score}} \tag{4.2}$$

avec S le nombre total de segments de parole et $T_{score} = \sum_{s=1}^{S} dur(s) \cdot N_{ref}$ le temps total sur le lequel le score est calculé. Les termes $N_{ref}(s)$ et $N_{sys}(s)$ donnent le nombre de locuteurs

2. cet outil d'évaluation est disponible sur http ://www.itl.nist.gov/iad/mig/tools/index.html

actifs dans le segment s. $N_{correct}(s)$ donne le nombre de locuteurs actifs proposés dans *sys* et correctement appariés dans *ref* pour le segment s. Les segments de non-parole contiennent par définition 0 locuteurs. L'equation (4.2) peut être décomposée en quatre types d'erreur dont la somme est le taux d'erreur de reconnaissance :

$$TER = E_{locuteur} + E_{fausse-alarme} + E_{non-détection} + E_{double-voix} \tag{4.3}$$

$E_{locuteur}$ est le pourcentage de temps attribué à un mauvais locuteur, sans tenir compte des phénomènes de double voix. $E_{fausse-alarme}$ tient compte des segments faussement étiquetés parole alors qu'au contaire, $E_{non-détection}$ comptabilise le temps des segments de parole non détectés. Enfin, $E_{double-voix}$ traduit le pourcentage de temps pour lesquels aucune des personnes qui parlent en même temps n'est reconnue. Il est parfois choisi de ne pas utiliser cette dernière erreur et d'écarter ainsi les segments de double-voix du calcul du taux d'erreur. Comme explicité par Anguera [2006], ces erreurs peuvent être détaillées de la façon suivante :

$$E_{locuteur} = \frac{\sum_{s=1}^{S} dur(s) \cdot \big(min\big(N_{ref}(s), N_{sys}(s)\big) - N_{correct}(s)\big)}{T_{score}} \tag{4.4}$$

$$E_{fausse-alarme} = \frac{\sum_{s=1}^{S} dur(s) \cdot \big(N_{sys}(s) - N_{ref}(s)\big)}{T_{score}} \qquad \forall \ \big(N_{sys}(s) - N_{ref}(s)\big) > 0 \tag{4.5}$$

$$E_{non-détection} = \frac{\sum_{s=1}^{S} dur(s) \cdot \big(N_{ref}(s) - N_{sys}(s)\big)}{T_{score}} \qquad \forall \ \big(N_{ref}(s) - N_{sys}(s)\big) > 0 \tag{4.6}$$

$E_{double-voix}$ se retrouve compris soit dans $E_{non-détection}$ soit dans $E_{fausse-alarme}$ suivant lesquels des fichiers de référence *ref* ou de sortie *sys* présentent des locuteurs non-assignés. L'erreur est comprise dans $E_{locuteur}$ si de multiples locuteurs apparaissent dans les fichiers de référence *ref* et de sortie *sys*. Lors de l'évaluation des scores de reconnaissance une tolérance peut être incluse sur la localisation temporelle de chaque tour de locuteur. Dans le cadre des campagnes NIST celle-ci est fixée à plus ou moins 0.25 secondes.

Cependant, comme le soulève Bonastre [2008], les taux d'erreur utilisés par la communauté ne peuvent pas être considérés comme des garants définitifs de la capacité de reconnaissance des systèmes. Poursuivant cette réflexion, nous avons remarqué que de telles mesures pouvaient dans certains cas ne pas refléter l'entièreté des capacités des algorithmes. Ainsi, le taux d'erreur proposé dans l'équation (4.3) est sensible à la répartition du temps de parole entre les locuteurs. Suivant les contenus à traiter — comme c'est le cas aux chapitres 5 et 6 — certains intervenants monopolisent plus la parole que d'autres. De fait, l'identification correcte de ces locuteurs majoritaires peut alors être suffisante à l'obtention de scores honorables.

À la suite de ce constat, nous avons proposé une première métrique alternative insensible au temps de parole de chaque locuteur (voir Vallet *et al.* [2010]) baptisée métrique ***unipond***. Celle-ci

est particulièrement bien adaptée aux émissions de talk show. En effet, comme nous l'avons vu dans le chapitre 1, deux personnes partagent généralement dans ces types d'émissions environ 70% du temps de parole : l'invité principal et le présentateur. Or, le but de ce travail de thèse étant la structuration de talk show, les interventions de tous les locuteurs sont potentiellement d'intérêt. Nous avons donc décidé d'inclure une pondération pour le calcul de $E_{locuteur_{unipond}}$:

$$E_{locuteur_{unipond}} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_{erreur}(i)}{T_{score}(i)} \tag{4.7}$$

$$TER_{unipond} = E_{locuteur_{unipond}} + E_{fausse-alarme} + E_{non-détection} \tag{4.8}$$

avec, pour N locuteurs, une fois l'identification locuteur/*cluster* effectuée, $T_{score}(i)$ la durée totale de l'intervention du locuteur i et $T_{erreur}(i)$ la durée totale mal attribuée pour ce même locuteur. Évidemment, une telle pondération attribuant à chaque intervenant un poids égal, sans distinction de volume de parole, peut sembler excessive. Cependant celle-ci permet de mettre en évidence certains écueils ne pouvant être observés avec les taux d'erreur de reconnaissance de type NIST. Sur le même principe nous avons proposé une autre pondération (métrique ***semi-pond***) présentée dans l'équation (4.9), plus réaliste qui équilibre à $\frac{50}{k}$% du temps total chacun des k locuteurs principaux et attribue de façon uniforme aux $N-k$ intervenants secondaires les 50% restant (dans le cas du *Grand Échiquier*, $k=2$) :

$$E_{locuteur_{semipond}} = \frac{1}{2}\left[\frac{1}{k}\sum_{i=1}^{k}\frac{T_{erreur}(i)}{T(i)} + \frac{1}{(N-k)}\sum_{j=k+1}^{N}\frac{T_{erreur}(j)}{T_{score}(j)}\right] \tag{4.9}$$

$$TER_{semipond} = E_{locuteur_{semipond}} + E_{fausse-alarme} + E_{non-détection} \tag{4.10}$$

Les métriques *unipond* et *semipond* ne prennant pas en compte les phénomènes de double-voix, nous utiliserons par la suite les résultats de taux d'erreur de reconnaissance donnés par la métrique NIST pour mesurer cet aspect.

4.6 Évaluation d'un système de reconnaissance de locuteurs issu de l'état de l'art

Après avoir effectué un tour d'horizon des techniques de reconnaissance de locuteurs, soulevé les spécificités propres au genre télévisuel du talk show et étudié les méthodes d'évaluation des algorithmes, nous proposons d'évaluer les résultats de reconnaissance obtenus avec un système issu de l'état de l'art. Celui exposé ici est basé sur l'algorithme de référence *top-down* proposé par Fredouille *et al.* [2009] et Bozonnet *et al.* [2010a].

4.6.1 Présentation de l'algorithme

Développé au LIA[3], le système utilisé se base sur l'utilisation de modèles de Markov à états cachés évolutifs (E-HMM, voir Meignier *et al.* [2001]) pour lesquels les états correspondent à des locuteurs et les transitions à des tours de parole. Les locuteurs sont modélisés par des mélanges de gaussiennes (GMM). Le système original est composé de quatre étapes successives comme cela est indiqué dans le tableau 4.2.

algorithme
Détection automatique de parole
boucle
Segmentation et regroupement
Purification (optionnel)
fin de boucle
Resegmentation

Tableau 4.2 – Étapes constitutives de l'algorithme de reconnaissance de locuteurs de Fredouille *et al.* [2009] et Bozonnet *et al.* [2010a].

Détection automatique de parole : cette étape consiste à identifier les segments de parole/non-parole de façon itérative au moyen d'un décodage de Viterbi sur un modèle de Markov à deux états cachés (un état pour la parole et l'autre pour la *non-parole*). Les états sont initialisés au moyen de modèles de mélanges de gaussiennes (GMM) et entrainés par l'algorithme Espérance-Maximisation (EM).

Segmentation et regroupement : des modèles sont isolés pour chaque locuteur de façon itérative. Le modèle de mélange de gaussiennes global est apris sur toutes les données de parole puis entrainé par l'algorithme Espérance-Maximisation (EM). Ce modèle est subdivisé à chaque itération et un nouveau modèle est appris. Ensuite une resegmentation des données est effectuée par décodage de Viterbi.

Purification : les regroupements effectués à l'étape précédente sont purifiés en utilisant les sous-segments qui sont les plus proches du modèle généré pour en apprendre un nouveau. Les sous-segments restants sont réassignés aux modèles les plus proches à la suite de plusieurs décodages de Viterbi (voir détails dans Bozonnet *et al.* [2010a]).

Resegmentation : les frontières des segments de parole sont redéfinies et certains segments, comme par exemple ceux des locuteurs parlant trop peu, sont éliminés comme à l'étape de segmentation.

4.6.2 Résultats

Nous exposons ici les résultats obtenus avec l'algorithme précédemment décrit dans deux contextes différents : celui des évaluation NIST RT'09 pour l'année 2009 et celui d'émissions de

3. Laboratoire Informatique d'Avignon - http ://lia.univ-avignon.fr/

talk show (corpus *Le Grand Échiquier*). Deux systèmes sont proposés : sans étape de purification (système *top-down TD* comme décrit dans Fredouille *et al.* [2009]) et avec étape de purification des *clusters* (système *top-down purification TDP*, voir Bozonnet *et al.* [2010a]).

4.6.2.1 corpus NIST RT

Le tableau 4.3 indique les scores obtenus par le système lors de la campagne NIST RT'09 pour la métrique NIST (avec et sans prise en compte du phénomène de double-voix). Il s'agit de résultats moyennés sur le corpus des sept émissions présentées dans la partie 4.4. Ces résultats attestent de la qualité de l'algorithme proposé, le plaçant parmis les meilleurs systèmes soumis pour l'évaluation 2009.

corpus	**système *TD***		**système *TDP***	
	avec double-voix	**sans double-voix**	**avec double-voix**	**sans double-voix**
NIST RT'09	26.0	21.5	21.1	16.0

Tableau 4.3 – Taux d'erreur NIST avec et sans évaluation du phénomène de double-voix pour la reconnaissance de locuteurs sur les 7 émissions du corpus NIST RT'09.

Il est intéressant d'apprécier l'amélioration résultant de l'ajout du module de purification des *clusters* proposé par Bozonnet *et al.* [2010a]. En effet, on observe une baisse des taux d'erreur avec et sans prise en compte du phénomène de double-voix de 5% en moyenne.

4.6.2.2 corpus *Le Grand Échiquier*

Les émissions du corpus NIST RT'09 ne comptent en moyenne que 13 minutes de parole ce qui est considérablement moins important que pour le corpus *Le Grand Échiquier* (voir tableau 4.1). Pour ce dernier, l'ensemble considéré de six émissions annotées en tours de parole a été divisé en deux pour obtenir un ensemble de développement et un ensemble de test de trois émissions chacun. Les paramètres de l'algorithme ont ainsi été ajustés au regard des performances obtenues sur l'ensemble de développement. De plus, le système n'étant pas construit pour traiter des données de grandes dimensions, la plus longue des émissions a été elle-même divisée en deux. Enfin, il est important de noter que l'étape de détection automatique de parole a été effectuée semi-manuellement en sélectionnant les séquences de parole supérieures à trente secondes à partir des annotations (ceci est détaillé plus finement dans le chapitre suivant).

Le système a obtenu les résultats décrits dans le tableau 4.4 pour la métrique NIST (avec et sans prise en compte du phénomène de double-voix). Le tableau 4.5 montre lui les résultats pour les métriques *unipond* et *semipond* décrites dans la partie 4.5 et mieux adaptées au genre télévisuel du talk show.

Pour le tableau 4.4, si les tendances générales semblent confirmées pour le corpus *Le Grand*

émission	**système *TD***		**système *TDP***	
	avec double-voix	**sans double-voix**	**avec double-voix**	**sans double-voix**
ensemble de développement				
CPB82055196	42.5	36.8	42.6	36.9
CPB84052346 (part. 1)	30.3	25.3	26.8	21.1
CPB84052346 (part. 2)	46.2	43.1	39.2	35.7
CPB85104049	44.4	38.4	45.4	38.6
moyenne	40.8	35.9	38.5	33.1
ensemble de test				
CPB82051110	45.6	43.1	43.0	40.3
CPB82051645	29.8	21.0	27.6	18.8
CPB88000401	52.1	51.2	52.6	51.8
moyenne	42.5	38.4	41.1	37.0

Tableau 4.4 – Taux d'erreur NIST pour la reconnaissance de locuteurs sur 6 émissions du *Grand Échiquier* pour les systèmes *TD* et *TDP* testés avec et sans prise en compte du phénomène de double-voix.

émission	**système *TD***		**système *TDP***	
	unipond	***semipond***	***unipond***	***semipond***
ensemble de développement				
CPB82055196	86.7	59.6	87.8	60.4
CPB84052346 (part. 1)	56.9	39.4	57.1	39.4
CPB84052346 (part. 2)	60.1	41.1	59.6	40.9
CPB85104049	73.7	49.6	70.7	47.5
moyenne	69.3	47.4	68.8	47.1
ensemble de test				
CPB82051110	65.2	45.4	62.1	43.4
CPB82051645	63.7	44.5	63.1	43.9
CPB88000401	75.7	48.6	75.9	49.6
moyenne	68.2	46.2	67.0	45.6

Tableau 4.5 – Taux d'erreur métriques *unipond* et *semipond* pour la reconnaissance de locuteurs sur 6 émissions du *Grand Échiquier* pour les systèmes *TD* et *TDP* testés sans prise en compte du phénomène de double-voix.

Échiquier avec une amélioration notable des scores lors de l'ajout de l'étape de purification, on peut observer que les performances sont très largement en deçà de celles obtenues lors de la campagne d'évaluation NIST. En effet, la décroissance observée est d'environ 15%. Pour les métriques *unipond* et *semipond* les taux d'erreur sont encore plus élevés (voir tableau 4.5) et par conséquent indiquent que nombre d'intervenants sont généralement mal distingués. Une analyse plus poussée a révélé que les très faibles scores de l'émission CPB82055196 étaient dûs à des conditions acoustiques particulièrement difficiles, des bruits parasites pouvant être entendus par intermittence sur la bande-son. De plus, cette émission comprend un nombre très important de locuteurs (vingt), ce qui a pour conséquence de faire chuter très fortement les résultats des métriques *unipond* et *semipond* si seulement une petite partie d'entre eux est correctement modélisée.

4.6.3 Conclusion de l'évaluation

Les résultats obtenus pour les six émissions du *Grand Échiquier* indiquent que les méthodes de reconnaissance de locuteurs issues de l'état de l'art se prêtent mal à l'étude de corpus de talk show. En effet dans ce dernier cas, les taux d'erreur augmentent drastiquement. Cela est d'autant plus visible à l'étude des scores fournis par les métriques *unipond* et *semipond*. Celles-ci ont en effet été créées pour répondre aux cas d'usage relatifs à un schéma de structuration de talk show. En effet dans ce cas, la détection correcte de locuteurs majoritaires, si elle assure de bonnes performances avec les métriques de type NIST, n'est pas suffisante dans une perspective d'organisation/indexation audiovisuelle. Les métriques *unipond* et *semipond* peuvent donc être vues comme des indicateurs relativement fiables de la qualité de reconnaissance des différents intervenants d'un plateau de talk show.

Conclusion

La reconnaissance de locuteurs est une thématique de recherche qui existe depuis plusieurs dizaines d'années. Cependant, c'est à partir des années 2000 que sont apparus des systèmes complètement non-supervisés. Auparavant, il fallait disposer de données d'apprentissage pour reconnaître un locuteur. Désormais, par reconnaissance de locuteurs (*speaker diarization*) on entend plutôt le découpage automatique d'un flux audio en segments homogènes labelés distinctement pour chaque participant, la correspondance entre labels et participants étant à effectuer ultérieurement par l'utilisateur.

Le problème de reconnaissance de locuteurs nécessite donc la résolution de plusieurs sous-problèmes. Généralement, il faut d'abord isoler les parties de parole dans le flux audio puis proposer une segmentation en tours de parole avant d'effectuer un regroupement hiérarchique audio de ceux-ci. La segmentation en tours de parole et le regroupement peuvent parfois être effectués de façon conjointe. La vaste majorité des sytèmes exploite exclusivement la modalité audio. Cependant, à la suite des travaux biométriques et avec la création de bases de données audiovisuelles, de nouveaux systèmes multimodaux ont également vu le jour.

L'étude comparative des caractéristiques des enregistrements de réunion de travail et des émissions de talk show a montré d'imporantes différences, notamment en ce qui concerne les durées à traiter et le nombre d'intervenants. Par conséquent, les méthodes d'évaluation des systèmes de reconnaissance de locuteurs, qui s'inscrivent généralement dans le cadre de campagnes d'évaluation telles que NIST ou ESTER, peuvent être sujettes à discussion. En effet, pour certains corpus, comme ceux constitués d'émissions de talk shows, on peut remettre en cause leur bien-fondé, les méthodes d'évaluation utilisées ne revèlant parfois qu'un point de vue biaisé et pouvant donner une vision incomplète des résultats de reconnaissance. C'est pour cette raison que nous avons proposé les métriques *unipond* et *semipond*, à priori plus sensibles à l'identification de tous les intervenants d'une émission et ce malgré les volumes de parole très variables.

Enfin, nous avons mesuré les performances d'un système de reconnaissance issu de l'état de l'art lorsqu'utilisé sur le corpus *Le Grand Échiquier*. Les résultats obtenus, au moyen des métriques NIST, *unipond* et *semipond* nous confortent dans l'idée que des améliorations doivent être apportées à la tâche de reconnaissance de locuteurs, l'identification des tours de parole étant cruciale pour l'élaboration du schéma de structuration que nous proposons.

Chapitre 5

Descripteurs visuels pour la reconnaissance de locuteurs

Introduction

Nous avons observé dans le chapitre 4 que les méthodes de reconnaissance de locuteurs proposées dans l'état de l'art étaient principalement basées sur l'analyse du seul flux audio. Cependant, très récemment, et ce en partie grâce à la constitution de nouvelles bases de données audiovisuelles, des méthodes multimodales ont été proposées pour la résolution de ce problème (en particulier Friedland *et al.* [2009a]). Dans ce chapitre, nous proposons d'extraire des caractéristiques du signal vidéo afin d'améliorer le système de reconnaissance de locuteurs de Fredouille *et al.* [2009] utilisé au chapitre précédent. En effet, l'étape d'initialisation des modèles de locuteurs est bien souvent critique pour ce genre de système, puisque les interventions d'un locuteur ne pourront être retrouvées dans le flux audio que si un modèle de parole lui correspondant est généré. Nous proposons donc d'utiliser des descripteurs vidéo pour identifier les différents intervenants et ainsi créer des modèles initiaux pour tous les locuteurs, même les moins représentés. Pour l'évaluation du système proposé nous utiliserons la métrique popularisée par les campagnes NIST RT ainsi que les deux nouvelles mesures proposées dans le chapitre précédent.

5.1 Extraction de caractéristiques audiovisuelles

Comme précédemment exposé, le talk show est un type de contenu télévisé particulier. En effet, contrairement aux bases de données utilisées en biométrie, le contenu vidéo est monté, c'est-à-dire que plusieurs caméras sont utilisées pour le tournage et qu'à chaque instant, les images d'une seule sont diffusées. Il existe bien entendu des cas particuliers de montage composés, comme par exemple la juxtaposition des visages de deux interlocuteurs filmés avec des caméras distinctes, mais ceux-ci ne représentent qu'une petite fraction des images montrées aux téléspectateurs sur l'ensemble d'un programme. Chaque plan est choisi par le réalisateur qui, généralement, essaie de suivre le locuteur à l'écran. Par conséquent, bien que le cadrage varie, (plan large, gros plan, plan américain, etc.), la plupart du temps « on voit qui l'on entend ». Cette observation motive l'intérêt de recourir à des descripteurs visuels bien que ceux-ci ne soient pas

d'une fiabilité absolue, le locuteur courant pouvant ne pas être montré à l'écran. Des descripteurs visuels sont donc extraits, comme cela est décrit plus bas, à partir des images des personnes à l'écran, en supposant qu'elles sont les locuteurs actifs. Il est important de mentionner que, les émissions des corpus de travail *Le Grand Échiquier* et *On n'a pas tout dit* étant réalisées en direct, notre tâche est compliquée par les conditions sonores bruyantes et la grande spontanéité de parole des intervenants.

5.1.1 Descripteurs audio

Le précédent chapitre a mis en évidence que dans la quasi-totalité des travaux de recherche en reconnaissance de locuteurs, les coefficients cepstraux à fréquence de Mel (MFCC) étaient les descripteurs les plus utilisés (voir Rabiner et Juang [1993]). Ceux-ci résultent du calcul d'une transformée en cosinus discrète appliquée au spectre de puissance d'un signal, les bandes de fréquence de ce spectre étant espacées logarithmiquement selon l'échelle perceptive de Mel. Ici également, le choix a été fait d'utiliser ces attributs audio. Pour cela, une fois la piste audio isolée comme cela a été décrit dans le chapitre 3, les MFCC (ainsi que leurs dérivées premières et secondes) ont été extraits en utilisant le logiciel *YAAFE*[1] de Mathieu *et al.* [2010].

5.1.2 Descripteurs video

Notre hypothèse étant qu' « on voit qui l'on entend », nous supposons que l'information portée par les torses des participants peut aider à l'identification des locuteurs.

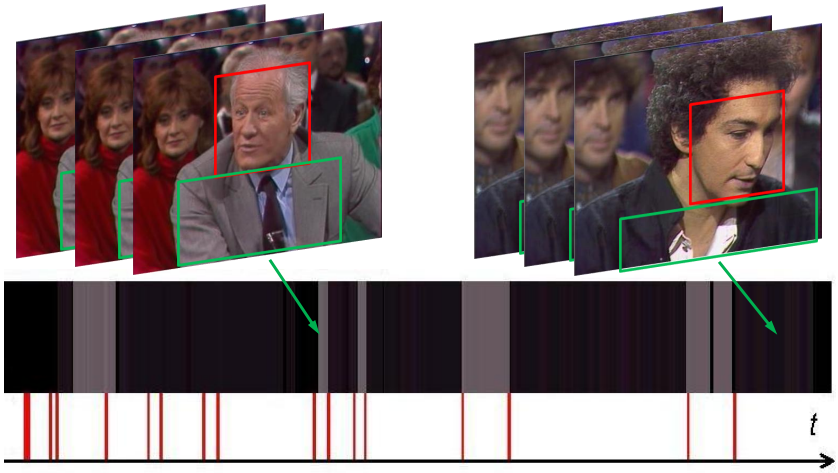


Figure 5.1 – Couleur dominante du costume et tours de parole (en rouge) pour un segment de parole de deux minutes de l'émission CPB85104049 du corpus *Le Grand Échiquier*.

1. Yet Another Audio Feature Extractor - http ://yaafe.sourceforge.net/

Celle-ci présente l'avantage de pouvoir être extraite de façon plus robuste que les attributs de visage. En effet, la plasticité de ces derniers, leur instabilité, leurs mouvements, les problèmes d'exposition à la lumière ainsi que le fait que de nombreux intervenants sont vus de profil font de la reconnaissance de visage une tâche très complexe. L'hypothèse simplificatrice que nous avançons ici est consolidée par l'importante corrélation observée entre couleur dominante du costume et tours de parole comme cela est illustré dans la figure 5.1.

Le schéma 5.2 donne une vue d'ensemble du processus d'extraction et de sélection des descripteurs audiovisuels.

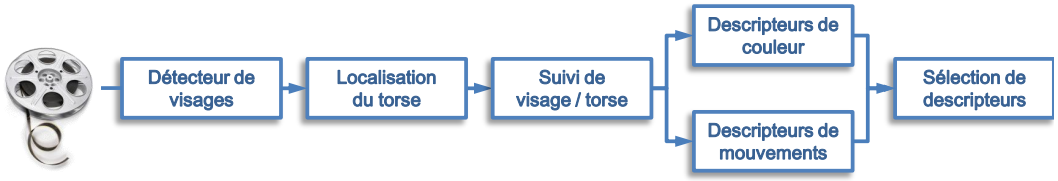


Figure 5.2 – Processus d'extraction et de sélection des descripteurs audiovisuels.

5.1.2.1 Détection de visage et localisation du torse

Inspirés par Jaffré et Joly [2004], Khoury *et al.* [2010] et plus généralement Jaffré [2005], nous proposons donc d'utiliser le costume comme attribut pour détecter automatiquement l'apparition d'une personne à l'écran. La présence d'un individu est attestée par la recherche de visages à l'intérieur de chaque trame. Nous utilisons pour cela l'algorithme de Viola et Jones [2001]. Une implémentation de celui-ci ainsi que des fichiers d'apprentissage sont disponibles dans la librarie *OpenCV*[2] (voir Bradski et Kaehler [2008]). Nous déterminons ensuite les régions des torses des intervenants en traçant des rectangles sous les visages détectés comme dans la thèse de Jaffré [2005]. Les dimensions du rectangle englobant le costume sont fixées comme suit :

$$H_{torse} = 5/2 \cdot H_{visage} \tag{5.1}$$

$$L_{torse} = 5/3 \cdot L_{visage} \tag{5.2}$$

avec H_k et L_k les hauteurs et largeurs des rectangles englobant visage et torse.

Détection de visage

La méthode de Viola et Jones pour la détection d'objets, et dans ce cas précis de visages, repose sur l'utilisation d'ondelettes similaires à celles de Haar comme attributs (voir figure 5.3).

2. Open Source Computer Vision - http ://sourceforge.net/projects/opencvlibrary/

En deux dimensions, il s'agit d'une combinaison de rectangles adjacents foncés et clairs. La présence ou l'absence d'une caractéristique pseudo-Haar est déterminée par la différence de la moyenne des pixels entre la zone foncée et la zone claire. Si cette différence est significative, c'est-à-dire au dessus d'un certain seuil, on en déduit la présence de cette caractéristique. Cela est effectué sur toute l'image et à différentes échelles très efficacement grâce à des astuces de programmation comme la méthode des images intégrales.

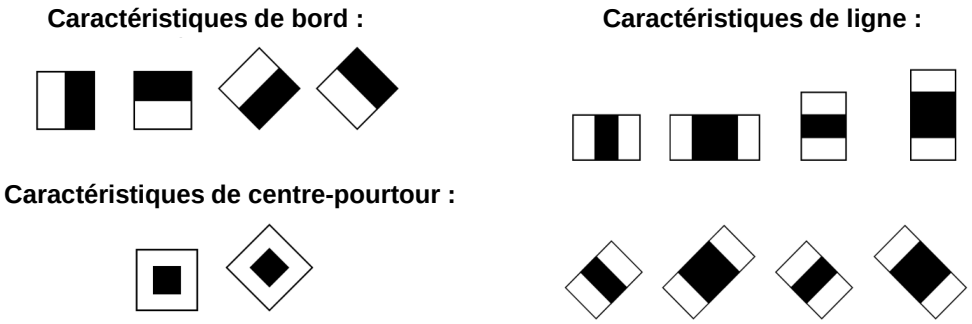


Figure 5.3 – Exemples de caractéristiques pseudo-Haar.

Ensuite, les caractéristiques sont sélectionnées en étant interprétées comme des ensembles de classifieurs faibles visage/non-visage qui peuvent être « boostés » par l'algorithme AdaBoost proposé par Freund et Schapire [1997] (voir également Freund et Schapire [1999]). La méthode propose une architecture pour combiner les classifieurs boostés en un processus en cascade, ce qui apporte un net gain en temps de détection, les classifieurs les plus simples étant présentés en premier pour rejeter la majorité des exemples négatifs très rapidement. Comme nous l'avons déjà souligné, l'étape d'entraînement pour la classification visage/non-visage est déjà réalisée dans la librarie *OpenCV*. Il est donc possible de procéder directement à la phase de test.

Dans notre cas, deux classifieurs sont utilisés pour détecter les visages : un de face et un de profil pour chaque image. De plus, nous limitons le nombre de détections dans chaque trame et indiquons une taille minimale pour les visages, pour ne garder finalement que la plus grande des régions d'intérêt proposées. Ce processus doit empêcher, autant que possible, la détection de visages dans le public (à l'arrière plan et donc souvent plus petits).

Suivi de visage

Cependant, la détection de visages trame à trame introduit de nombreuses fausses alarmes et non-détections. Pour s'affranchir de ce problème nous utilisons une heuristique simple exploitant les propriétés temporelles de la vidéo. Ainsi, nous proposons de faire l'économie d'un suivi de visage puisqu'une analyse de flux optique est effectuée par la suite pour l'extraction de descripteurs de mouvement et nous exploitons les propriétés de cohérence spatiale et temporelle à l'intérieur d'un plan. La figure 5.4 résume le procédé d'extraction des régions d'intérêt.

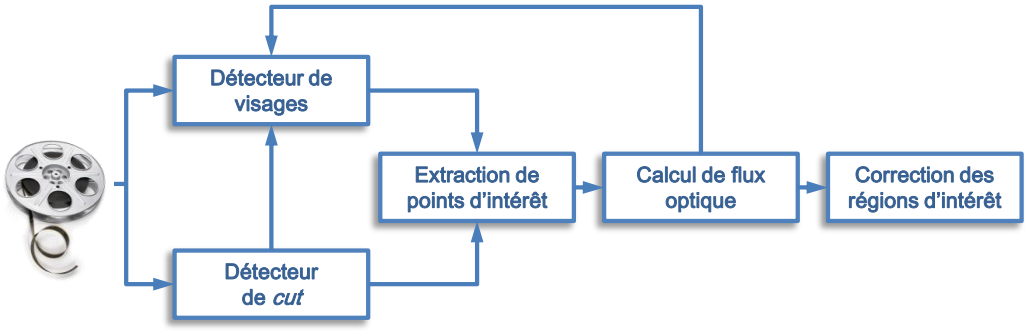


Figure 5.4 – Procédé d'extraction des régions d'intérêt de visage et de torse.

Après avoir implémenté le détecteur de changement de plans *Shotdetect*[3] du projet *ADVENE*[4] du LIRIS[5] (voir Aubert et Prié [2007]), nous extrayons des points d'intérêt sur les rectangles contenant visage et torse avec l'algorithme de Shi et Tomasi [1994]. Ensuite, ces points d'intérêt sont suivis à l'intérieur d'un plan grâce à l'algorithme de Lucas et Kanade [1981] (voir annexe G). Nous appelerons $t_{début}$ et t_{fin} les temps de début et de fin de ce plan. Le suivi est initialisé avec un nombre maximum de 300 points d'intérêt au temps t_{max} correspondant à la trame contenant le visage de plus grande taille à l'intérieur du plan.

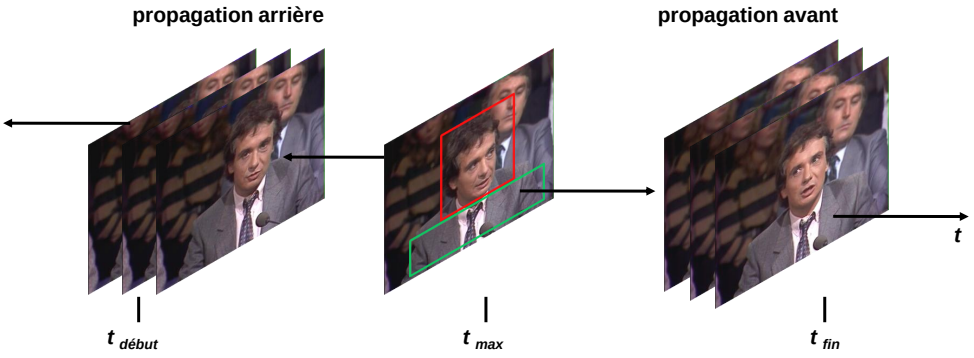


Figure 5.5 – Suivi de visage et torse à l'intérieur d'un plan.

Le suivi est ensuite propagé en avant jusqu'au temps t_{fin} et en arrière jusqu'au temps $t_{début}$, soit à la fin et au début du plan courant (voir figure 5.5). On peut également noter que dans le cas où plus d'un tiers des points suivis sont perdus entre deux trames (indiquant généralement la présence d'un changement de plan non-détecté) le suivi est arrêté (voir figure 5.6). Cette pro-

3. Shotdetect - http ://shotdetect.nonutc.fr/
4. Annotate Digital Video, Exchange on the Net - http ://liris.cnrs.fr/advene/
5. Laboratoire d'InfoRmatique en Image et Systèmes d'information - http ://liris.cnrs.fr/

cédure est répétée tout le long de l'émission et les erreurs de détection de visage sont corrigées entre $t_{début}$ et t_{fin} de la manière suivante. Si une trame f ne présente pas de visages détectés entre $t_{début}$ et t_{fin} les régions d'intérêt du visage et du torse de la trame $f-1$ (ou $f+1$ dans le cas de la propagation arrière) sont utilisées pour déduire celles manquantes par translation (déduites du déplacement des points d'intérêt suivis). De plus, si entre $t_{début}$ et t_{fin} le déplacement entre les trames f et $f+1$ (ou f et $f-1$ dans le cas de la propagation arrière) des régions d'intérêt est trop important par rapport à celui des points d'intérêt, nous supposons que le visage détecté à $f+1$ (respectivement $f-1$) est différent de celui détecté à f. Il correspond souvent à un visage au second plan qui peut ponctuellement devenir prépondérant pour la trame $f+1$ (respectivement $f-1$) à cause de variations du cadrage. Par conséquent, les régions d'intérêt correspondantes pour le mauvais visage et le mauvais torse sont supprimées et d'autres valides sont créées par translation de celles de la trame f comme expliqué précédemment.

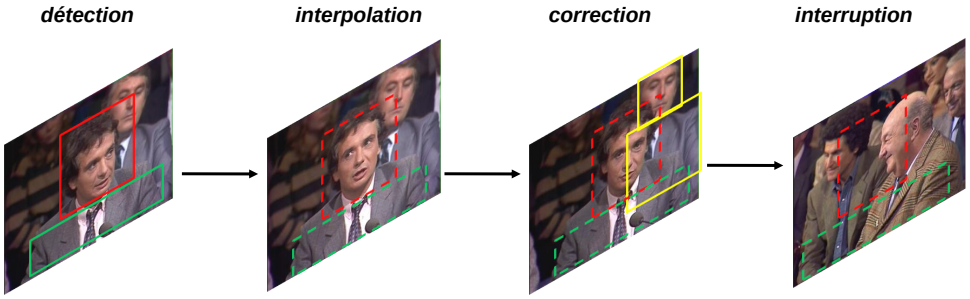


Figure 5.6 – Suivi de visage et torse à l'intérieur d'un plan avec les différents scénarii possibles : interpolation, correction (si le visage détecté dans la trame traitée est trop éloigné du visage interpolé), interruption (en cas de changement de plan non détecté).

5.1.2.2 Extraction de descripteurs de couleur

Une fois obtenue une description robuste des occurrences des torses des personnes à l'écran dans une vidéo, nous proposons de calculer un attribut sur les régions d'intérêt correspondantes reposant sur le MPEG-7 Dominant Color Descriptor (voir Manjunath *et al.* [2002]). Celui-ci propose une description compacte de 1 à 8 couleurs dominantes pour une image ou région d'intérêt, une couleur dominante étant un vecteur de trois composantes RGB $[u_R, u_G, u_B]$, chacune quantifiée sur 32 bins. Le descripteur calculé est la couleur dominante moyenne qui est une moyenne pondérée des d couleurs recouvrant au moins 40% des pixels de la région d'intérêt selon :

$$u_{C_{avg}} = \frac{\sum_{i=1}^{d} P_i . u_{C_i}}{\sum_{i=1}^{d} P_i} \tag{5.3}$$

avec

$$\sum_{i=1}^{d} P_i \geq 40\% \tag{5.4}$$

u_{C_i} étant la valeur de la i$^{\text{ème}}$ couleur dominante et P_i la proportion de pixels de cette couleur. Ce descripteur témoigne d'une signature colorimétrique caractéristique de costume. Le ratio de 40% s'avère être une estimation robuste de la surface couverte par le costume dans la région d'intérêt, en prenant en compte les conditions potentiellement bruitées survenant par exemple lorsque les mains du locuteur entrent dans le rectangle englobant le torse.

Des histogrammes de couleur HSV (*Hue Saturation Value*) sur les régions d'intérêt des costumes sont également extraits avec respectivement 16, 4 et 2 bins pour des amplitudes variant de 0 à 180 pour la composante H et de 0 à 255 pour S et V.

5.1.2.3 Extraction de descripteurs de mouvement

Inspiré de Friedland *et al.* [2009a], nous proposons aussi de calculer plusieurs descripteurs de mouvement basés sur l'analyse du flux optique. Nous supposons que chaque locuteur possède ses propres gestuelles et expressions qui, décrites avec les bons attributs, peuvent être très discriminantes. On peut par exemple penser au mouvement des mains souvent visible dans la région d'intérêt du torse. Pour caractériser ces particularités de façon robuste et efficace, nous proposons de déduire du flux optique des descripteurs de mouvement pour les deux régions conjointement puis de facon différenciée comme montré dans la Figure 5.7.

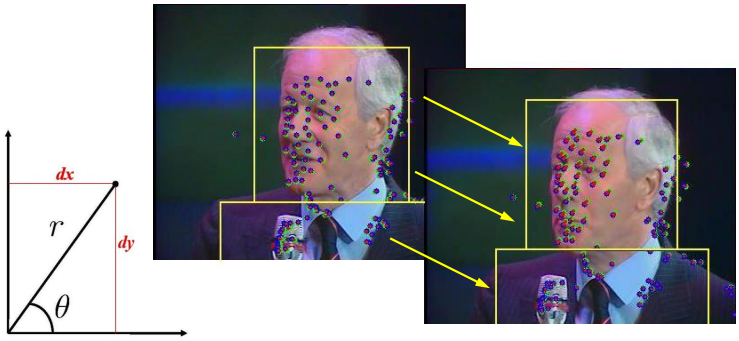


Figure 5.7 – Mouvements entre deux trames consécutives pour une sélection de points d'intérêt. Le repère situé à gauche explicite les amplitudes de mouvement r et orientation θ pour un déplacement (dx,dy).

Les amplitudes et orientations de vitesse et d'accéleration sont calculées comme les dérivées première et seconde des points d'intérêt de l'image globale et des régions d'intérêt, celles-ci étant évaluées comme les coordonnées r et θ dans un système polaire. Nous proposons également de calculer l'amplitude relative présentée comme le rapport des amplitudes pour les régions d'intérêt du visage et de la poitrine sur celle de l'image globale.

Nous avons également proposé le calcul d'un histogramme d'orientations de mouvements de 36 bins (soit une sensibilité de 10 degrés) pour les points d'intérêt de la région du torse. Un histogramme pondéré par l'amplitude de mouvement de chaque point a aussi été calculé.

5.1.2.4 Extraction de descripteurs globaux MPEG-7

Un grand nombre de descripteurs vidéo globaux issus de l'état de l'art ont été extraits. Nous avons en particulier utilisé les attributs MPEG-7 (voir Manjunath *et al.* [2002]) calculés à l'aide du logiciel XM (*eXperimentation Model*) développé par le groupe de travail *Moving Picture Experts Group* (MPEG). Le tableau 5.1 présente les descripteurs globaux extraits des vidéos.

famille	descripteur	dimension	détail
couleur	*ColorLayout*	12	Distribution spatiale des couleurs
	ColorStructure	32	Structure locale de la couleur
	ImageHistogram	48	Histogramme dans l'espace YUV
	ScalableColor	64	Quantification de l'espace HSV
texture	*EdgeHistogram*	80	Distribution locale des contours
	HomogeneousTextureGabor	24	Texture par filtres de Gabor
	HomogeneousTextureHaralick	78	Texture par filtres de Haralick
forme	*RegionShape*	35	Distribution des pixels

Tableau 5.1 – Descripteurs vidéo globaux MPEG-7 extraits.

5.1.2.5 Sélection automatique de descripteurs

Afin de retenir les caractéristiques les plus pertinentes pour notre problème parmi celles ayant été initialement extraites, nous exploitons un système de sélection automatique des descripteurs (*feature selection*, voir Guyon et Elisseeff [2003]). Le critère de sélection choisi pour cela est un ratio inertie inter-classes sur inertie intra-classe s'apparentant à un ratio de Fisher (voir Duda *et al.* [2000]). Celui-ci indique une mesure du degré de séparabilité des classes de locuteurs avec chaque ensemble de descripteurs considéré. Plus celui-ci est élevé et plus les classes sont bien distinguées. Par conséquent, cette étape est effectuée de façon supervisée sur une partie de l'émission annotée CPB85104049 de la base de développement du corpus *Le Grand Échiquier*. Il est calculé suivant :

$$J = \frac{\sum_{i=1}^{c} n_i (\tilde{m}_i - \tilde{m})^t (\tilde{m}_i - \tilde{m})}{\sum_{i=1}^{c} \sum_{x \in C_i} (x - \tilde{m}_i)^t (x - \tilde{m}_i)} \tag{5.5}$$

avec

$$\tilde{m}_i = \frac{1}{n_i} \sum_{x \in C_i} x \qquad \text{et} \qquad \tilde{m} = \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{m}_i$$

c étant le nombre de *clusters*, $x \in C_i$ les vecteurs de caractéristiques du *cluster* i avec n_i leur

nombre et $\tilde{m}_i$ le centroïde du *cluster*. De plus, n est le nombre total de vecteurs et $\tilde{m}$ le centroïde pour tous les *clusters*. La figure 5.8 propose une visualisation des scores de séparabilité obtenus.

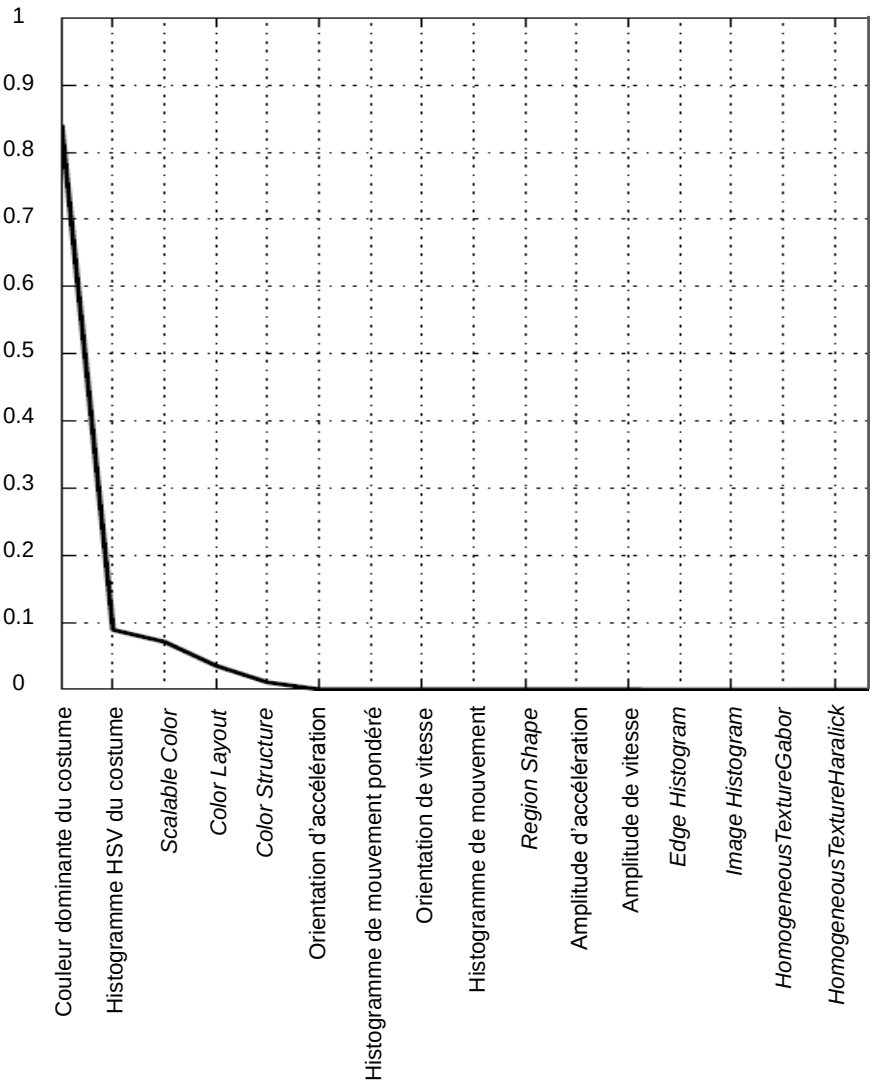


Figure 5.8 – Valeurs de séparabilité J pour la sélection automatique des descripteurs visuels.

Les descripteurs les plus discriminants sont ceux pour lesquels la séparation inter-classe, mesurée par la distance entre les vecteurs de caractéristiques de chaque cluster et les autres centroïdes, est maximale et la séparation intra-classe, mesurée pour chaque cluster par la distance des vecteurs au centroïde, minimale. Les descripteurs pour lesquels le ratio est maximum sont les caractéristiques de couleur dominante de costume. Notre objet d'étude étant du contenu télévisé, ce n'est pas à proprement parler surprenant. En effet, il est souvent demandé aux participants d'un talk show de porter des couleurs distinctives afin que le téléspectateur ait plus

de facilité à les distinguer. Par conséquent, ces descripteurs sont utilisés dans l'algorithme de reconnaissance de locuteurs *AV1* décrit par la suite.

5.2 Initialisation audiovisuelle d'un système : algorithme *AV1*

Ce travail a été l'objet d'une collaboration avec Simon Bozonnet (EURECOM [6]) dans le cadre du projet SELIA [7] dirigé par Nicholas Evans (EURECOM).

Nous avons focalisé notre travail sur l'amélioration de l'étape d'initialisation d'un système de reconnaissance de locuteurs (*speaker diarization*) issu de l'état de l'art, en l'occurence celui proposé par Fredouille *et al.* [2009] et Bozonnet *et al.* [2010a]. Celui-ci est présenté dans le chapitre précédent à l'occasion de l'étude des performances d'un système état de l'art pour des émissions de talk show. Le système de Bozonnet *et al.* [2010a] se distingue de celui de Fredouille *et al.* [2009] par l'ajout d'une étape de purification des modèles de locuteurs (expliquée chapitre 4). Rappelons très succintement les quatre étapes de cet algorithme (voir tableau 5.2).

algorithme
Détection automatique de parole
boucle
Segmentation et regroupement
Purification (optionnel)
fin de boucle
Resegmentation

Tableau 5.2 – Étapes constitutives de l'algorithme de reconnaissance de locuteurs de base.

Notons également que la première étape est effectuée ici aussi de façon semi-automatique en sélectionnant les sections de parole de plus de trente secondes à partir des annotations manuelles afin de mettre l'accent sur la tâche de reconnaissance proprement dite. Nous utilisons le terme de section de parole pour désigner par la suite chaque bloc de parole précédant et suivant des éléments de contenus longs de plus de dix secondes (voir figure 5.9).

L'étape d'initialisation des systèmes de reconnaissance, qu'ils soient agglomératifs ou divisifs, est particulièrement importante puisque si aucun modèle n'est proposé pour un locuteur, celui-ci ne pourra pas être retrouvé par la suite. Dans la version que nous proposons ici, l'étape de segmentation et de regroupement est initialisée par un *pré-clustering* effectué sur les descripteurs visuels de couleur dominante de costume présentés précédemment. Nous nous basons ainsi sur l'hypothèse que pour les programmes de talk show « on voit qui l'on entend ». Cela est en effet motivé par les choix de montage effectués par le réalisateur de l'émission. Les descripteurs visuels sont beaucoup plus stables temporellement que ceux extraits du signal audio, très sensibles aux changements de conditions acoustiques. Il est donc possible, grâce à l'hypo-

6. EURECOM - http ://www.eurecom.fr/

7. Suivi et séparation de Locuteurs pour l'Indexation Audio - projet interne à l'Institut Télécom (collaboration EURECOM / Télécom ParisTech)

thèse que nous avons avancée quant au montage des émissions de talk show, de proposer de façon non-supervisée des segments d'apprentissage pour les différents locuteurs identifiés par leur costume dans l'émission. Les autres étapes de l'algorithme de reconnaissance de locuteurs restent ensuite inchangées.

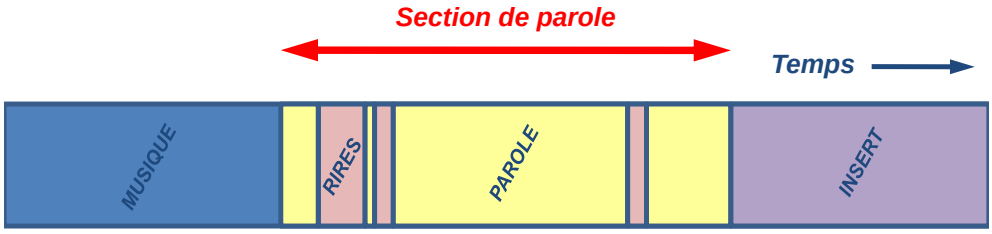


Figure 5.9 – Schéma illustratif d'une *section* de parole.

Nous formulons l'hypothèse que chaque section de parole contient entre deux et dix locuteurs. Un *clustering* par k-moyennes partitionnant les données visuelles de chaque section en K groupes est ensuite utilisé en faisant varier K entre 2 et 10 (voir Duda *et al.* [2000]). Les données visuelles utilisées sont les couleurs dominantes de costume présentées dans la section précédente. Pour chaque K testé, 100 classifications différentes sont effectuées et on conserve celle pour laquelle la somme des distances à chaque centroïde est minimale. Ensuite, nous utilisons la méthode proposée par Sugar et James [2003] et basée sur un calcul de « distorsion » d_K pour déterminer le nombre de partitions K à conserver :

$$d_K = \frac{1}{l} \min_{c_1,\ldots,c_K} E\left[(x-c_x)^T \Sigma^{-1} (x-c_x)\right] \tag{5.6}$$

avec x un vecteur de caractéristiques de dimension l, Σ la matrice de covariance, K le nombre de *clusters* testés et $(c_1, c_2, \ldots, c_K)$ les centroïdes des *clusters*. La distorsion est la distance de Mahalanobis moyenne de chaque point au centroïde le plus proche. Le bon nombre de partitions est déterminé par le calcul de « sauts » :

$$S_K = d_K^{-Y} - d_{K-1}^{-Y} \tag{5.7}$$

avec Y une transformation en puissance ($Y > 0$). Le saut le plus important est observé à la suite d'une amélioration brutale des résultats qui s'explique par la séparation en deux d'une partition contenant deux classes différentes. L'ajout de groupes supplémentaires réduit ensuite davantage la distorsion intra-groupe que la distorsion inter-groupe et a donc un impact moindre. Par conséquent, le nombre correct de *clusters* est sélectionné comme celui présentant le plus grand « saut ». La transformation en puissance choisie est celle proposée par Sugar et James [2003] : $Y = l/2$. La figure 5.10 indique l'allure des courbes de « sauts ».

Seuls les *clusters* de plus de dix secondes sont ensuite conservés, les données des autres étant réassignées aux *clusters* restants. De nouveaux modèles acoustiques sont alors appris sur les

données correspondants aux partitions visuelles obtenues et l'étape de segmentation et de regroupement se poursuit ensuite comme dans Fredouille *et al.* [2009] et Bozonnet *et al.* [2010a].

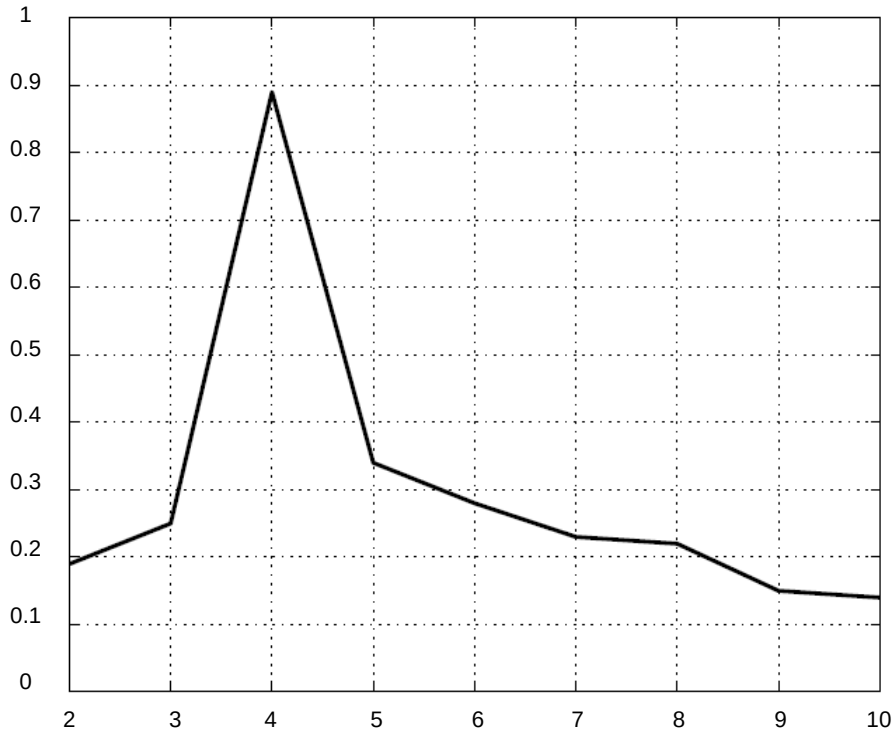


Figure 5.10 – Valeurs des « sauts » pour $K \in [2,10]$ pour une section de parole de l'émission CPB85104049 du *Grand Échiquier*. La valeur choisie pour le nombre de clusters est $K = 4$.

5.3 Résultats et discussion

Nous avons déjà évoqué la qualité de l'algorithme de base utilisé ici dans le chapitre 4. Ainsi, nous avons pu noter qu'il avait obtenu de très bons résultats lors de la campagne d'évaluation NIST RT'09. Cependant, appliqué au type de contenu particulier que sont les émissions de talk show, ce système offre des performances contrastées, qu'il s'agisse de la version *TD* (*top-down*, voir Fredouille *et al.* [2009]) ou de la version *TDP* (*top-down purification*, voir Bozonnet *et al.* [2010a]).

L'étape d'initialisation audiovisuelle que nous avons précédemment présentée a été implémentée sur les deux versions *TD* et *TDP* donnant ainsi naissance aux systèmes *TDAV* (*top-down audiovisual*) et *TDPAV* (*top-down purification audiovisual*). Les résultats présentés dans le tableau 5.3 ont été obtenus avec ces quatre systèmes pour la métrique utilisée dans la campagne d'évaluation NIST. Par conséquent, les résultats des deux premières colonnes sont identiques à ceux présentés dans le chapitre précédent.

Là encore, en comparant les résultats des tableaux 4.3 du chapitre précédent et 5.3 on peut noter que les résultats obtenus sur le corpus *Le Grand Échiquier* sont — comme on pouvait s'y attendre en raison des conditions acoustiques, du grand nombre de locuteurs et de la durée des émissions — largement en deçà des performances obtenues lors de la campagne d'évaluation NIST RT'09. Cependant, on peut noter une amélioration visible des performances lorsque les descripteurs visuels sont utilisés pour l'initialisation du système de base. En moyenne, les taux d'erreur passent ainsi pour le système *top-down* de 40.8%/35.9% à 33.9%/29.9% pour la base de développement, et de 42.5%/38.4% à 35.9%/32.8% pour la base de test. De même, une décroissance similaire des taux d'erreur est observée lorsqu'on adjoint l'étape de purification des modèles de locuteurs au système de base : de 38.5%/33.1% à 30.0%/25.5% pour la base de développement et de 41.1%/37.0% à 35.2%/32.1% pour la base de test. Les résultats moyens obtenus sur la base de test sont marginalement moins bons que ceux de la base de développement. Cet aspect reflète les différences de difficulté qui peuvent être observées pour diverses émissions de talk show. En effet, les résultats sont fortement dépendants du temps de parole à traiter, du nombre de locuteurs, de la proportion de double-voix, etc. Enfin, pour certaines émissions comme CPB84052346 (part. 2) ou CPB85104049, on peut observer des décroissances assez spectaculaires en passant des systèmes 1 à 4. Celles-ci témoignent de la création d'au moins un nouveau modèle adapté à un locuteur alors que ce dernier était auparavant regroupé avec un autre intervenant.

émission	**système *TD***	**système *TDP***	**système *TDAV***	**système *TDPAV***
ensemble de développement				
CPB82055196	42.5 - 36.8	42.6 - 36.9	48.3 - 46.0	48.5 - 46.2
CPB84052346 (part. 1)	30.3 - 25.3	26.8 - 21.1	23.9 - 20.1	23.9 - 20.1
CPB84052346 (part. 2)	46.2 - 43.1	39.2 - 35.7	39.0 - 35.9	23.2 - 17.9
CPB85104049	44.4 - 38.4	45.4 - 38.6	24.5 - 17.5	24.6 - 17.6
moyenne	40.8 - 35.9	38.5 - 33.1	33.9 - 29.9	30.0 - 25.5
ensemble de test				
CPB82051110	45.6 - 43.1	43.0 - 40.3	33.6 - 31.2	34.5 - 32.0
CPB82051645	29.8 - 21.0	27.6 - 18.8	25.2 - 18.5	25.2 - 18.5
CPB88000401	52.1 - 51.2	52.6 - 51.8	48.9 - 48.6	46.1 - 45.7
moyenne	42.5 - 38.4	41.1 - 37.0	35.9 - 32.8	35.2 - 32.1

Tableau 5.3 – Taux d'erreur NIST pour la reconnaissance de locuteurs sur 6 émissions du *Grand Échiquier* pour les quatre systèmes testés avec et sans prise en compte du phénomène de double-voix.

Ayant étudié la répartition de la parole dans les émissions de talk show au chapitre 1, nous avons observé que certains locuteurs intervenaient beaucoup plus que d'autres. Or, la plupart des mesures utilisées en reconnaissance de locuteurs étant sensibles à la proportion de parole correctement attribuée, l'identification correcte des deux locuteurs principaux suffit dans le cas du talk show pour obtenir un taux d'erreur de reconnaissance honorable. En effet, ceux-ci se partagent en moyenne entre 60% et 70% du temps de parole total. Dans la première partie de cette thèse nous avons insisté sur l'importance de l'identification des tours de parole et sur la connaissance de leurs auteurs. En effet, les tours de parole sont considérés comme l'entité fondamentale

de structuration dans le schéma que nous proposons et l'identification d'un maximum d'intervenants est essentielle pour sa réalisation. Les deux nouvelles mesures *unipond* et *semipond* présentées au chapitre 4 ont ainsi été créées afin de refléter la capacité des algorithmes de reconnaissance à identifier correctement tous les intervenants d'une émission, chose qui n'est pas directement observable avec les taux d'erreur de reconnaissance de type NIST.

émission	**système *TD***	**système *TDP***	**système *TDAV***	**système *TDPAV***
ensemble de développement				
CPB82055196	86.7	87.8	92.6	92.8
CPB84052346 (part. 1)	56.9	57.1	54.0	54.0
CPB84052346 (part. 2)	60.1	59.6	52.7	49.9
CPB85104049	73.7	70.7	33.7	33.6
moyenne	69.3	68.8	58.2	57.6
ensemble de test				
CPB82051110	65.2	62.1	46.0	46.1
CPB82051645	63.7	63.1	49.3	53.8
CPB88000401	75.7	75.9	74.4	71.4
moyenne	68.2	67.0	56.5	57.1

Tableau 5.4 – Taux d'erreur métrique *unipond* pour la reconnaissance de locuteurs sur 6 émissions du *Grand Échiquier* pour les quatre systèmes testés. Ces mesures ne prennent pas en compte le phénomène de double-voix.

Le calcul de la première de ces métriques (*unipond*) donne les résultats présentés dans le tableau 5.4. Comme nous l'avons expliqué précédemment, celle-ci attribue un poids équivalent à tous les locuteurs, indépendamment de leur temps de parole respectif. Ainsi que nous l'avons laissé entendre lors de sa présentation au chapitre 4, les résultats observés avec cette mesure sont largement inférieurs à ceux obtenus avec le taux d'erreur de reconnaissance NIST. Cela confirme en effet l'intuition selon laquelle seuls les locuteurs les plus actifs — et par conséquent pour lesquels la création de modèles de parole est la plus aisée à réaliser — sont correctement identifiés. Là encore, il est intéressant de noter que l'utilisation des descripteurs visuels pour l'initialisation offre de notables améliorations par rapport aux systèmes *TD* et *TDP* uniquement audio. De plus, on peut remarquer que plusieurs des émissions testées se distinguent par leurs taux de reconnaissance extrême. D'un côté, l'émission CPB82055196 présente de très mauvais résultats. Cela est dû à des conditions acoustiques particulièrement difficiles, des bruits parasites pouvant être entendus par intermittence sur la bande-son. Il est notable que c'est le seul cas pour lequel l'initialisation visuelle détériore les résultats. De plus, cette émission comprend vingt locuteurs, ce qui a pour conséquence de faire chuter drastiquement les résultats si un nombre important d'entre eux ne sont pas correctement modélisés. D'un autre côté, l'émission CPB85104049 montre des scores très respectables. Ceux-ci peuvent en partie être expliqués par le nombre de locuteurs actifs au cours de l'émission (dix), sensiblement moins que dans le reste des émissions du corpus *Le Grand Échiquier* (comme cela est indiqué dans le chapitre 1).

Il va sans dire que la métrique *unipond*, si elle est informative, peut être considérée comme trop radicale. En effet, on peut débattre de l'intérêt d'attribuer un poids équivalent à des indi-

vidus parlant de quelques secondes à un plus d'un quart d'heure. Ainsi, ayant remarqué que généralement les temps de parole du présentateur et de l'invité principal représentent entre 60% et 70% du temps de parole total lors d'un talk show, nous avons proposé une seconde métrique (*semipond*) attribuant 50% du score de reconnaissance à ces deux locuteurs majoritaires et répartissant les 50% restants entre les autres intervenants, avec des poids équivalents, comme pour la métrique *unipond* (voir la description au chapitre 4). Les résultats obtenus sont présentés dans le tableau 5.5. En raison de l'équilibrage artificiel effectué entre locuteurs principaux et secondaires, les taux d'erreur se situent à mi-chemin entre ceux obtenus avec la métrique NIST et la métrique *unipond*. Les mêmes phénomènes sont observés que dans les tableaux 5.3 et 5.4 avec les meilleurs scores obtenus pour le système *TDPAV*.

émission	**système *TD***	**système *TDP***	**système *TDAV***	**système *TDPAV***
ensemble de développement				
CPB82055196	59.6	60.4	66.4	66.6
CPB84052346 (part. 1)	39.4	39.4	39.3	39.3
CPB84052346 (part. 2)	41.1	40.9	46.5	36.5
CPB85104049	49.6	47.5	23.9	23.9
moyenne	47.4	47.1	44.0	41.6
ensemble de test				
CPB82051110	45.4	43.4	33.1	33.1
CPB82051645	44.5	43.9	37.0	39.7
CPB88000401	48.6	49.6	55.4	52.4
moyenne	46.2	45.6	41.8	41.7

Tableau 5.5 – Taux d'erreur métrique *semipond* pour la reconnaissance de locuteurs sur 6 émissions du *Grand Échiquier* pour les quatre systèmes testés. Ces mesures ne prennent pas en compte le phénomène de double-voix.

Enfin, il est intéressant de remarquer que les métriques *unipond* et *semipond* mettent toutes deux en évidence l'apport de l'utilisation de descripteurs visuels pour l'initialisation du système de reconnaissance plus que le taux d'erreur de reconnaissance NIST. En effet, alors que le système *TDP* améliore légèrement les résultats du système *TD* et qu'il en est généralement de même pour les systèmes *TDPAV* et *TDAV* (exceptions faites des émissions CPB82055196 et CPB88000401), des décroissances beaucoup plus fortes sont observées lors du passage à l'initialisation vidéo. Celles-ci n'étaient pas visibles de façon aussi évidente dans le tableau 5.3 et nous confortent dans l'idée d'avoir proposé de nouvelles métriques pour l'évaluation de systèmes de reconnaissance de locuteurs afin d'appréhender de façon plus compréhensive les résultats fournis par les algorithmes.

5.4 Exploitation des descripteurs audiovisuels en phase de classification

Les résultats précédemment présentés ont prouvé l'utilité des descripteurs audiovisuels pour l'initialisation d'un système de reconnaissance de locuteurs. Cependant, on est en droit de se demander quels auraient été les résultats obtenus si, à l'instar de Friedland *et al.* [2009c], des descripteurs visuels avaient été incorporés directement au processus de classification des trames de parole (étape de segmentation/regroupement). En collaboration avec Simon Bozonnet (EURECOM), des tentatives ont été effectuées dans ce sens mais celles-ci n'ont malheureusement pas abouties à de très probants résultats.

Ayant ainsi identifié une des limites du schéma de classification de l'algorithme *AV1*, nous avons proposé une nouvelle approche de reconnaissance de locuteurs visant à utiliser conjointement des descripteurs visuels et audio lors de l'étape de classification (voir Vallet *et al.* [2010]). Ainsi, nous avons considéré les descripteurs visuels de couleur et de mouvement calculés sur la région d'intérêt du torse en plus des traditionnels coefficients MFCC. La méthode que nous avons employée peut être qualifiée de semi-automatique puisque des exemples d'apprentissage pour chaque locuteur ont été fournis en amont par un utilisateur (un documentaliste de l'Ina par exemple) afin d'entraîner un classifieur SVM. Ainsi, un unique segment par locuteur, d'une durée comprise entre quatre et quinze secondes, a été fourni au classifieur (plusieurs tirages ont été effectués pour s'assurer de la validité de la méthode).

Les résultats obtenus avec ce nouveau système ont indiqué l'intérêt d'utiliser conjointement des descripteurs visuels et audio lors de l'étape de classification. De plus, ils ont mis en évidence le fait que le choix d'exemples d'apprentissage très fiables, même en très faible quantité, est critique et que si cette condition est remplie, les méthodes à noyau de type SVM peuvent fournir des outils de classification très performants.

Conclusion

Ayant mis en évidence précédemment l'importance de la reconnaissance de locuteurs pour l'élaboration d'un schéma de structuration pour les programmes de talk show ainsi que les limites des sytèmes état de l'art pour ce même type de contenu, nous avons proposé dans ce chapitre l'amélioration d'un système de base par utilisation de caractéristiques visuelles. Ainsi, en nous basant sur l'algorithme de Fredouille *et al.* [2009] et Bozonnet *et al.* [2010a], nous avons introduit une étape d'initialisation fondée sur l'information visuelle véhiculée par les costumes portés par les intervenants (en l'occurence la couleur dominante). En effet, dans un contexte télévisuel, les tenues des participants sont généralement soigneusement choisies afin que les téléspectateurs puissent aisément les identifier.

Les résultats de cette étude ont confirmé l'intérêt d'utiliser l'information visuelle pour la tâche de reconnaissance lorsque celle-ci est disponible, comme c'est le cas avec des émissions

de télévision. En effet, une amélioration des scores est observable que l'évaluation soit effectuée avec la mesure traditionnellement utilisée lors des campagnes NIST ou avec les deux nouvelles métriques *unipond* et *semipond* (détaillées dans le chapitre 4). Celles-ci sont un indicateur particulièrement précieux pour appréhender la proportion des intervenants d'une émission de talk show correctement identifiés.

Enfin, nous avons mis en évidence dans une autre étude que les descripteurs visuels pouvaient être précieux, non seulement pour la phase d'initialisation de systèmes de reconnaissance de locuteurs, mais également pour la classification des trames de parole. Nous avons par ailleurs pu noter à l'occasion de ce travail que de satisfaisants résultats pouvaient être obtenus par l'utilisation conjointe de techniques de classification type SVM et d'un nombre très limité d'exemples d'apprentissage d'une grande fiabilité.

Chapitre 6

Architecture de reconnaissance multimodale de locuteurs

Introduction

Au regard des encourageants résultats de reconnaissance de locuteurs obtenus lors de l'utilisation de descripteurs visuels, nous proposons d'élaborer un nouveau système de reconnaissance de locuteurs lui aussi complètement non-supervisé. Nous concentrons particulièrement nos efforts sur la constitution de modèles de parole fiables pour les différents intervenants dont l'importance a été soulevée au chapitre précédent. Pour cela nous utilisons à nouveau les données colorimétriques véhiculées par les costumes des personnes à l'écran. De plus, nous nous assurons que ces personnes sont effectivement dans un acte de conversation en calculant un taux d'activité labiale indicatif du mouvement des lèvres. Ayant constaté les bonnes performances pouvant être obtenues à l'aide de méthodes à noyau (voir Vallet *et al.* [2010]), nous procédons, une fois les données d'apprentissage obtenues, à une classification par machine à vecteurs de support (SVM) des caractéristiques audio extraites. Nous proposons ensuite de recourir à un classifieur SVM vidéo parallèle afin d'utiliser ce type de descripteurs non seulement pour l'initialisation du système mais également pour la phase de classification. En effet, les données visuelles présentent l'avantage d'être plus stables au cours du temps (en particulier à l'intérieur d'un plan) que leurs homologues audio. Afin de choisir laquelle des deux sorties de classifieur proposées doit être considérée, nous calculons une mesure de cohérence audiovisuelle (à l'aide de SVM à une classe). Si la cohérence entre les deux sorties est forte, le classifieur vidéo est choisi. À l'inverse, si elle est faible, c'est le classifieur audio qui est choisi. Nous mesurons ensuite les performances de notre système sur les corpus *Le Grand Échiquier* et *On n'a pas tout dit*, les métriques utilisées étant à nouveau celles de la campagne d'évaluation NIST et *unipond* et *semipond* présentées au chapitre 4.

6.1 Exploitation de méthodes à noyau : algorithme *AV2*

Le chapitre 5 a été l'occasion de présenter une première approche exploitant la modalité visuelle de façon relativement limitée et basée sur l'utilisation conjointe d'informations audio et vidéo dans un contexte de reconnaissance de locuteurs. À la lecture de travaux récents issus de la

littérature scientifique, il semble en effet que la corrélation des deux modalités puisse apporter de précieuses indications sur l'identité des intervenants au cours d'émissions télévisuelles.

Suite à une première étude (voir Vallet *et al.* [2010]), nous avons pu constater que d'encourageants résultats de reconnaissance de locuteurs pouvaient être obtenus en utilisant des méthodes de classification par machines à vecteurs de support (SVM). De plus, nous avons observé que dans ce cas, même avec très peu de données d'apprentissage, si celles-ci sont suffisamment fiables, des taux de reconnaissance satisfaisants pouvaient être atteints. C'est à la suite de ces réflexions que nous avons élaboré l'algorithme non-supervisé présenté dans la figure 6.1 et le tableau 6.1.

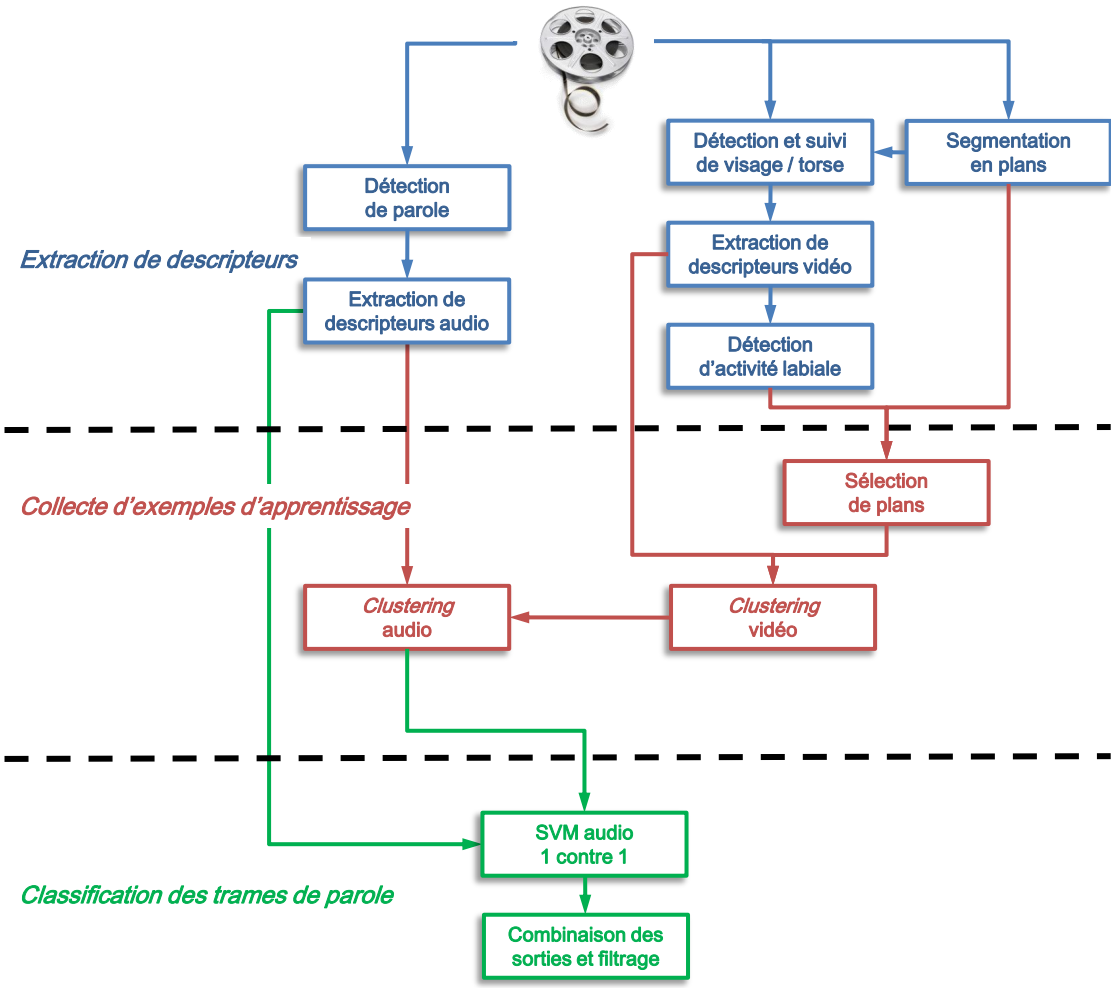


Figure 6.1 – Présentation de l'algorithme *AV2*.

On peut remarquer que dans l'algorithme proposé une part importante est accordée à la préparation de la base d'apprentissage. En effet, nous cherchons ainsi à obtenir de façon non-supervisée, et pour chaque locuteur, des exemples d'apprentissage fiables.

étapes			
1	-	détection semi-automatique de parole	
2	-	extraction de descripteurs audio	
3	-	détection de changements de plans	
4	-	détection de visages	*extraction de descripteurs*
5	-	correction de regions d'intérêt	
6	-	extraction de descripteurs vidéo	
7	-	détection d'activité labiale	
8	-	sélection de plans	
9	-	*clustering* vidéo (distance d_{χ^2})	*collecte d'exemples d'apprentissage*
10	-	*clustering* audio (distance probabiliste en RKHS)	
11	-	classification audio SVM (1 contre 1)	
12	-	évaluation des probabilités de sortie audio	*classification des trames de parole*
13	-	filtrage médian	

Tableau 6.1 – Présentation des différentes étapes constitutives de l'algorithme *AV2*.

Comme cela est détaillé au chapitre 5, la détection des segments de parole est effectuée de manière semi-automatique en utilisant les annotations manuelles afin de se concentrer ici sur la tâche de reconnaissance.

6.1.1 Extraction des descripteurs audio et vidéo

Les techniques d'extraction de descripteurs utilisées ici sont similaires à celles présentées au chapitre 5. Cependant, il existe des différences au niveau des descripteurs extraits. Ainsi, pour la modalité audio les coefficients LSF (*Line Spectral Frequency*, voir Schussler [1976] et Backstrom et Magi [2006]) sont ajoutés aux MFCC (et à leurs dérivées premières et secondes) précédemment calculés afin d'enrichir la paramétrisation du système et de garantir plus de robustesse. Pour la modalité vidéo, nous conservons le suivi de visage et de torse et extrayons les mêmes descripteurs de mouvement qu'au chapitre 5. En revanche, les seules caractéristiques colorimétriques calculées sont des histogrammes de couleur HSV de costume ainsi que des versions cumulées de ces histogrammes pour chaque plan pour lequel un costume est présent (comme au chapitre 5). Le tableau 6.2 synthétise les descripteurs extraits et donne leurs dimensions.

modalité	**descripteur**	**dimension**	**détail**
audio	*MFCC*, Δ, Δ^2	39	Coefficients cepstraux et dérivées
	LSF	10	Coefficients de prédicition linéaire
vidéo	*Hist HSV*	22	Histogrammes HSV de costume
	HistCum HSV	22	Histogrammes HSV de costume cumulés (par plan)

Tableau 6.2 – Descripteurs utilisés dans cette étude.

6.1.2 Détection de l'activité labiale

Le changement principal pour l'étape d'extraction de caractéristiques réside dans l'ajout d'un détecteur d'activité labiale. Inspiré par les travaux de Everingham *et al.* [2006] et Sivic *et al.* [2009], nous proposons ainsi de mesurer les amplitudes des mouvements qui peuvent être observés dans la région de la bouche des individus montrés à l'écran. Pour cela, nous sélectionnons les plans contenant les trames les plus sûres, c'est-à-dire celles pour lesquelles une détection de visage de face a été observée. Nous utilisons ensuite la méthode proposée par Everingham *et al.* [2006] pour localiser les commissures des lèvres sur les visages sélectionnés (et du même coup les autres points d'intérêt du visage : les yeux et le nez).

Pour localiser les neuf points caractéristiques de visage (coins droit et gauche de chaque oeil, narines et extrémité du nez et commissures des lèvres) un modèle d'apparence discriminant est calculé. Celui-ci exploite l'algorithme AdaBoost et des caractéristiques pseudo-Haar ((voir Viola et Jones [2001]). Un modèle génératif de la position des points d'intérêt est couplé au classifieur précédent. Il permet de fixer des contraintes de dépendance entre les points en modélisant leur distribution conjointe de façon analogue à celle proposée par Felzenszwalb et Huttenlocher [2005]. Les paramètres des modèles d'apparence et de forme des points caractéristiques ont été fixés par apprentissage à partir de données étiquetées.

Une fois les commissures des lèvres localisées pour les trames sélectionnées, une grille de points d'intérêt placés tous les deux pixels est disposée à l'intérieur d'un rectangle englobant la bouche. Ce rectangle est décalé légèrement vers le bas par rapport à l'axe défini par les commissures, la lèvre inférieure étant celle qui bouge le plus lors d'une prise de parole. Un calcul de flux optique de Lucas et Kanade est ensuite effectué sur toute la durée du plan, comme cela était le cas pour le suivi de visage. Là encore, suivant la position des trames de visage sélectionnées, la propagation peut être soit avant soit arrière.

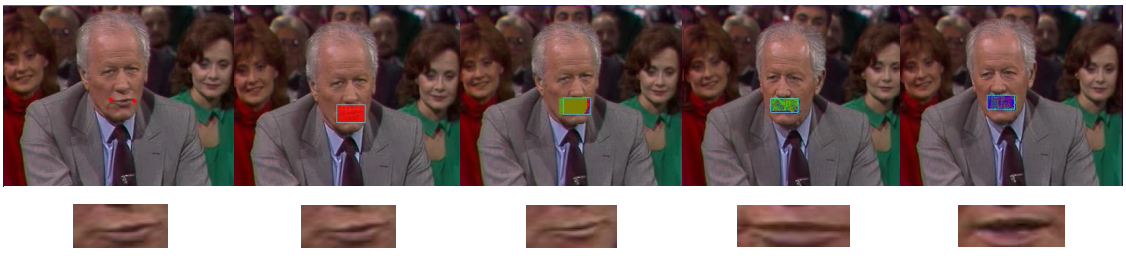


Figure 6.2 – Détection des lèvres et suivi des points d'intérêt dans la région de la bouche.

Cette méthode est illustrée par la figure 6.2. De gauche à droite peuvent ainsi être observés : un exemple de détection de bouche, la disposition d'une grille de points d'intérêt dans une boîte englobante puis le suivi par calcul de flux optique de ceux-ci pour trois trames. Les images des lèvres comprises dans les rectangles englobants sont affichées en dessous des trames correspondantes. Les points rouges sur la seconde image sont les points d'intérêt placés tous les deux pixels dans le rectangle englobant la bouche. Ces points d'intérêt observés à la trame $f-1$

apparaissent en vert, ceux observés à la trame f en bleu et leur déplacement est visualisable en rouge. L'activité labiale peut ensuite être calculée pour chaque trame f comme la différence du mouvement global de la tête (mesuré lors de l'extraction de descripteurs de mouvement) et de celle de la bouche. Ces mouvements globaux sont les moyennes des amplitudes de mouvement des points suivis dans les régions d'intérêt comme cela est présenté dans le chapitre 5. La figure 6.3 illustre le détecteur d'activité labiale ainsi implémenté. Les nombreuses variations observées sont ensuite nivelées par l'ajout d'un filtre médian d'une durée d'une demi-seconde et recouvrant à 50%. Cependant, ne disposant pas d'annotation visuelle en « visage parlant », nous n'avons pas pu évaluer la contribution de ce détecteur de façon directe. La validité de celui-ci a donc été établie a posteriori par les résultats de reconnaissance de locuteurs obtenus.

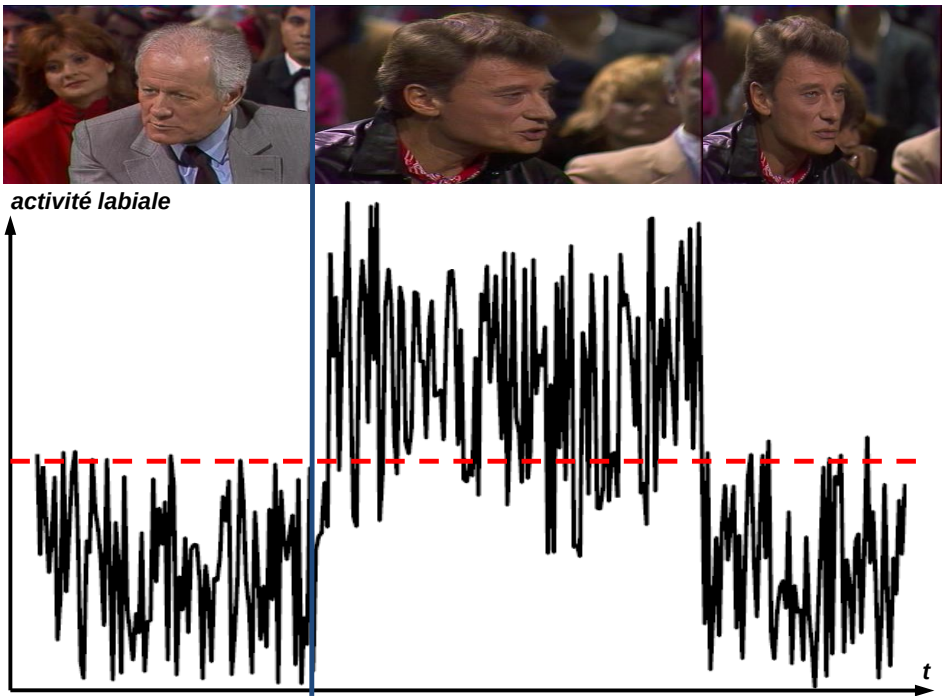


Figure 6.3 – Détection de l'activité labiale pour deux plans consécutifs. Le premier plan montre Jacques Chancel écoutant Johnny Halliday qui parle hors caméra. Le second montre Johnny Halliday parlant puis écoutant une intervention de Jacques Chancel.

Grâce aux nombreux points d'intérêt suivis au cours de chaque plan, ce détecteur d'activité labiale présente l'avantage d'être plus robuste pour la tâche considérée que celui proposé par Everingham *et al.* [2006]. En effet, dans ce dernier cas les auteurs testent leur algorithme de détection sur le corpus de séries télévisées *Buffy contre les vampires*. Or, le contenu d'émissions de talk show et celui de sitcoms est très différent. Les sitcoms mettent en scènes des acteurs suivant un scénario et étant dirigés par un réalisateur. Par conséquent, les mouvements erratiques qui peuvent être observés dans les programmes de talk show et au cours desquels les intervenants sont susceptibles de beaucoup bouger, d'effectuer des mouvements brusques, de subir

des changements soudain d'illumination, etc. sont généralement absents des séries télévisées. La facture de celles-ci est plutôt à rapprocher du document cinématographique.

6.1.3 Collecte d'exemples d'apprentissage

L'algorithme présenté au chapitre précédent a mis en évidence deux aspects essentiels : l'importance de l'étape d'initialisation pour les systèmes de reconnaissance de locuteurs et l'intérêt d'utiliser la composante visuelle lors du traitement de documents vidéo (et en particulier des émissions de talk show). Par conséquent, un effort important a été apporté à l'amélioration du processus de sélection non-supervisé des données d'apprentissage de modèles de locuteurs.

6.1.3.1 Sélection de plans

Conservant l'hypothèse selon laquelle, lorsque le contenu à traiter est de nature télévisuelle, le réalisateur montre la plupart du temps le locuteur actif, nous proposons une méthode de sélection des plans pertinents. Ceci dans le but de ne conserver que ceux pour lesquels un locuteur est à l'écran tout du long et de collecter ainsi des exemple d'apprentissage très fiables. Ces plans sont sélectionnés s'ils ont une longueur supérieure à un seuil fixé préalablement. Selon le type de réalisation, celui-ci peut varier. Par exemple, pour les émissions du corpus *Le Grand Échiquier*, un seuil relativement long peut être fixé (dix secondes). En revanche, pour celles d'*On n'a pas tout dit* pour lesquelles les échanges sont beaucoup plus brefs et succincts, il doit être abaissé (cinq secondes). Ces valeurs ont été choisies en rapport avec la durée moyenne d'un plan pendant les séquences parlées des émissions de l'ensemble de développement : 7.5 secondes pour *Le Grand Échiquier* contre 3 secondes pour *On n'a pas tout dit*. De plus, les plans sélectionnés pour l'apprentissage doivent présenter des locuteurs à l'écran. Parmi les plans d'une longueur suffisante sont donc conservés ceux pour lesquels un visage est détecté et une activité labiale moyenne importante est observée. Le seuil d'activité labiale est fixé également à l'aide des émissions constituant la base de développement

6.1.3.2 *Clustering* vidéo

Les histogrammes cumulés de couleur de costume ayant nécessairement été extraits pour les plans ainsi sélectionnés (puisqu'ils contiennent des visages), nous proposons d'effectuer une classification par regroupement hiérarchique ou *clustering*. Celui-ci permet ainsi de regrouper entre eux les plans de contenus visuels (de costume) similaires. Pour cela sont agrégés entre eux les plans de distance minimale, c'est à dire les plus proches. Ceux-ci constituent alors un nouveau *cluster* et le regroupement continue jusqu'à ce qu'il n'en reste qu'un seul. La figure 6.4 propose une illustration d'un tel *clustering*.

Nous avons utilisé la distance d_{χ^2} qui est connue pour s'appliquer particulièrement bien aux données de type histogramme. Celle-ci est issu du fameux test du χ^2 (voir Snedecor et Cochran [1967]).

Pour deux histogrammes Ha et Hb de b bins chacun nous avons ainsi :

$$d_{\chi^2} = \frac{1}{2}\sum_{i=1}^{b}\frac{(Ha_i - Hb_i)^2}{(Ha_i + Hb_i)} \tag{6.1}$$

Choisir un critère d'arrêt à une méthode de regroupement hiérarchique est toujours délicat. Afin d'éviter l'agrégation de plans qui pourraient être trop hétérogènes et ainsi contenir les données de plus d'un locuteur, nous choisissons d'arrêter le *clustering* largement au-dessus du nombre de locuteurs attendus. En effet, plus un *cluster* contient de données dissemblables et plus il aura tendance, par gravité, à en attirer de nouvelles. Dans le système proposé nous fixons arbitrairement à quarante le nombre de *clusters*, soit au moins deux fois plus que de locuteurs véritables.

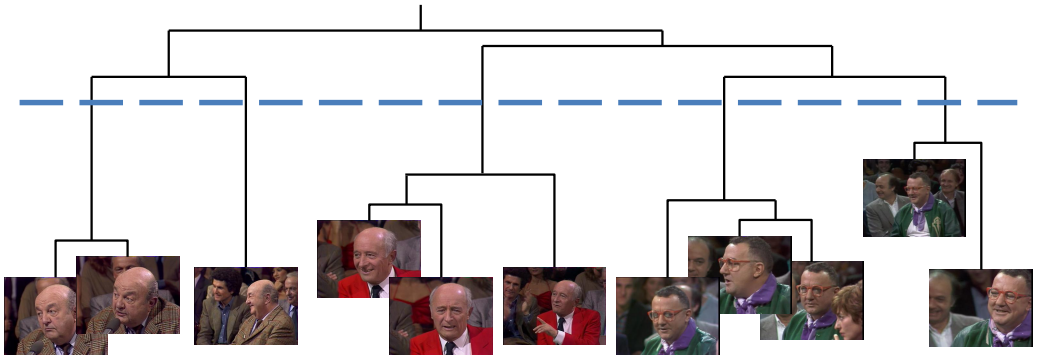


Figure 6.4 – Dendrogramme illustrant le regroupement hiérarchique de plans suivant les histogrammes de couleur cumulés du costume. Le trait pointillé bleu indique l'arrêt du *clustering* lorsque le critère est atteint.

6.1.3.3 *Clustering* audio

Bien que de nombreuses précautions aient été prises quant à la sélection des plans (constitués de locuteurs à l'écran) et à leur regroupement, certaines erreurs sont néanmoins susceptibles de survenir lors du *clustering* hiérarchique vidéo. Il convient donc d'utiliser la modalité audio pour obtenir les ensembles d'apprentissage des locuteurs finaux. De plus, comme nous l'avons déjà souligné, les descripteurs audio de type MFCC et LSF présentent beaucoup moins de stabilité au cours du temps que leurs homologues visuels (et en particulier colorimétriques).

Une fois le *clustering* hiérarchique visuel effectué, les partitions (*clusters*) obtenues sont de tailles importantes. Comme cela sera indiqué dans le tableau 6.3, on passe en moyenne de plus d'une centaine de plans sélectionnés longs d'au moins dix secondes chacun à quarante *clusters*. Il devient dès lors beaucoup plus facile de procéder à un traitement audio puisque beaucoup plus de données sont à disposition pour chaque *cluster*. Il est ainsi possible d'apprendre

des distributions audio pour les descripteurs MFCC (ainsi que leurs dérivées premières et secondes) et LSF extraits précédemment. Ceci est d'ailleurs renforcé par la différence entre les taux d'échantillonnage pour l'extraction des descripteurs vidéo et audio (25Hz et 100Hz) qui fait correspondre pour un descripteur vidéo quatre descripteurs audio.

Par conséquent, nous proposons de regrouper les partitions obtenues précédemment en calculant les distances probabilistes dans un espace de Hilbert à noyau reproductible (RKHS) entre chaque paire de *clusters* supposés distribués dans cet espace suivant des lois gaussiennes. Comme proposé par Essid [2005], nous utilisons la distance de Bhattacharryya pour mesurer la dissimilarité des distributions. L'expression générale de celle-ci est donnée dans l'équation (6.2) pour deux densités de probabilités p et q :

$$d_{BC}(p,q) = -\log\left(\int_u \sqrt{p(u)q(u)}\,du\right) \tag{6.2}$$

Si les densités de probabilités sont considérées comme gaussiennes, la distance de Bhattacharryya peut alors être calculée selon :

$$d_{BC}(p,q) = \frac{1}{8}(\mu_p-\mu_q)^T\left[\frac{1}{2}(\Sigma_p+\Sigma_q)\right]^{-1}(\mu_p-\mu_q) + \frac{1}{2}\log\frac{\left[\frac{1}{2}(\Sigma_p+\Sigma_q)\right]}{\left[\Sigma_p\right]^{\frac{1}{2}}\left[\Sigma_q\right]^{\frac{1}{2}}} \tag{6.3}$$

Le principe d'utilisation est ensuite d'estimer les paramètres des gaussiennes $N(\mu_p,\Sigma_p)$ et $N(\mu_q,\Sigma_q)$ dans l'espace de Hilbert à noyau reproductible (RKHS) en utilisant l'astuce du noyau (*kernel trick*). Plus de détails sur l'utilisation de distances probabilistes en RKHS peuvent être trouvés dans les travaux de Zhou et Chellappa [2006]. Est ainsi obtenu une matrice des distances inter-partitions. L'appariement des *clusters* se fait par regroupement hiérarchique. Cependant, contrairement à l'étape précédente, le critère d'arrêt du regroupement est fixé à l'aide d'un seuil (et non d'un nombre de partitions attendu). Ce seuil est déterminé suivant le corpus traité sur la base développement en déterminant empiriquement le meilleur moment pour arrêter le *clustering*.

6.1.3.4 Évaluation de la collecte d'exemples d'apprentissage

Il est possible d'évaluer la pertinence du procédé de sélection des exemples proposé. Le tableau 6.3 indique l'évolution de la pureté des *clusters* tout au long des étapes de sélection de plans, *clustering* vidéo et *clustering* audio ainsi que la durée de l'ensemble d'apprentissage ainsi obtenu. Cette pureté est calculée comme une précision à savoir qu'elle est égale, pour chaque *cluster*, au ratio du temps de parole correctement attribué au locuteur considéré sur le temps de parole attribué à ce même locuteur.

L'évolution des scores de pureté indique que notre méthode de sélection d'exemples est robuste. En effet, au cours du processus le nombre de *clusters* est significativement abaissé sans que pour autant une perte en pureté soit observée, permettant d'aboutir ainsi à la création d'un nombre limité de modèles de parole très purs pour les intervenants des émissions.

émission	**sélection de plans**		***clustering* vidéo**		***clustering* audio**		**durée**	**nb. locuteurs**
	nb. *clusters*	pureté	nb. *clusters*	pureté	nb. *clusters*	pureté		
CPB82055196	78	87.6%	40	82.6%	16	81.6%	30'28"	20
CPB84052346	148	93.3%	40	88.6%	15	88.3%	41'06"	15
CPB85104049	96	94.3%	40	93.2%	10	93.1%	29'09"	10
moyenne	107.3	91.7%	40	88.1%	13.7	87.7%	33'34"	15

Tableau 6.3 – Évolution de la pureté et du nombre de *clusters* au cours du processus de constitution de la base d'apprentissage pour les émissions de l'ensemble d'entraînement du corpus *Le Grand Échiquier*. La durée indique le temps total représenté par les exemples sélectionnés.

6.1.4 Classification des trames de parole

À l'issue de la sélection des exemples d'apprentissage, il est possible d'effectuer une classification des trames identifiées lors de la détection semi-automatique de parole. Pour cela nous choisissons d'utiliser des classifieurs SVM un contre un avec les descripteurs audio MFCC (avec dérivées premières et secondes) et LSF, ceux-ci ayant déjà donné de bons résultats dans un travail présenté antérieurement (voir Vallet *et al.* [2010]).

Nous proposons d'utiliser les machines à vecteurs de support (SVM) présentées dans l'annexe H pour la classification des trames de parole suivant les classes proposées de façon non-supervisée à l'issue de la collecte d'exemples d'apprentissage.

6.1.4.1 Problèmes biclasses

Dans notre étude, les partitions issues de la phase de constitution de la base d'apprentissage permettent de créer des classifieurs SVM un contre un. Ainsi, pour chaque problème biclasse un hyperplan séparateur peut être calculé. Par conséquent, pour un ensemble d'apprentissage de N *clusters* il y a $\frac{N(N-1)}{2}$ problèmes biclasses à résoudre. Les facteurs de coût C_+ et C_- sont définis comme étant le rapport du nombre d'exemples positifs sur celui d'exemples négatifs. Enfin, le noyau utilisé est le noyau gaussien défini suivant :

$$k(x,y) = \exp\left(-\frac{||x-y||^2}{2l\sigma_k^2}\right) \tag{6.4}$$

avec l la dimension du vecteur de caractéristiques x. Le paramètre σ_k est lui fixé pour toutes les expériences par validation croisée sur les émissions de l'ensemble de développement. Pour la tâche de classification SVM nous avons utilisé la boîte à outils (*toolbox*) *LIBSVM*[1] (voir Chang et Lin [2011]).

1. http ://www.csie.ntu.edu.tw/~cjlin/libsvm/

6.1.4.2 Combinaison des sorties de classifieurs

Une fois obtenue pour chaque échantillon de parole de la base de test et chaque classifieur une valeur pour la fonction g présentée à l'équation (H.6) de l'annexe H, nous combinons ces sorties dans le but d'obtenir pour chaque vecteur d'observation des probabilités d'appartenance pour les N *clusters* appris. On passe alors de $\frac{N(N-1)}{2}$ problèmes biclasses à un unique problème multiclasse. Afin d'obtenir des sorties probabilistes pour chacun des $\frac{N(N-1)}{2}$ classifieurs, nous utilisons l'approche de Platt [1999]. Les probabilités a posteriori $P(y=1|g)$, f étant donné par l'équation (H.6), sont alors modélisées par :

$$P(y=1|g) = \frac{1}{1+\exp(Af+B)} \tag{6.5}$$

où A et B sont des paramètres à déterminer. Platt montre la pertinence de ce modèle et propose un algorithme pour la détermination des valeurs optimales de A et B. Il devient ensuite possible de fusionner les sorties probabilistes grâce au couplage par paire (dit *minpair*) proposé par Wu *et al.* [2004]. Celui-ci est implémenté dans une extension de la boîte à outils *LIBSVM*.

6.1.4.3 Filtrage

Une fois les probabilités multiclasses obtenues, un filtrage median est effectué. Pour cela une fenêtre glissante de taille 0.5 seconde (soit deux fois la tolérance fixée dans les campagnes d'évaluation NIST) et recouvrante à 50% est utilisée. Sur chacune de ces fenêtres on attribue au segment temporel correspondant l'étiquette (*label*) pour laquelle la probabilité d'un locuteur est maximale.

6.2 Résultats et discussion

Comme pour les résultats présentés au chapitre précédent, nous proposons une évaluation de la reconnaissance de locuteurs à l'aide de trois métriques différentes. Un avantage du système que nous proposons ici est qu'il est relativement insensible à la longueur des émissions à traiter. Ainsi, nous donnons des scores globaux pour l'émission CPB84052346 ainsi que sur chacune des deux parties comme cela est fait au chapitre 5. L'intérêt de traiter une émission d'un bloc est que les étiquettes proposées ne nécessitent d'être identifiées au locuteur correspondant qu'une fois pour toutes. De plus, il est possible de construire des ensembles d'apprentissage de taille plus importante. Le tableau 6.4 donne les résultats du système proposé pour les trois métriques.

En comparant ces résultats avec ceux obtenus au chapitre 5 on peut remarquer que, sur la métrique NIST, le système *TDPAV* de l'algorithme *AV1* reste en moyenne plus efficace que l'algorithme *AV2* sur la base de développement : 30.0%/25.5% contre 34.5%/30.5%. Ce n'est par contre plus le cas sur la base de test : 35.2%/32.1% contre 33.6%/30.6%. Pour les deux autres métriques (*unipond* et *semipond*) en revanche, le nouveau système présente de meilleurs résultats. En effet, avec la métrique *unipond*, les taux d'erreur passent de 57.6% à 38.6% pour la

base de développement et de 56.5% à 54.2% pour la base de test. Cette dernière amélioration est nettement moins importante que celle observée sur la base de développement. De façon similaire pour la métrique *semipond* il passe de 41.6% à 33.4% et reste à 41.7% pour la base de test. Ces différences de résultats semblent indiquer que plus de locuteurs sont identifiés avec l'algorithme *AV2* mais que les intervenants majoritaires sont légèrement moins bien définis. De plus, si les résultats obtenus avec la métrique NIST sont très similaires sur les bases de développement et de test, on peut remarquer que pour les deux autres mesures, ils sont en moyenne plus dégradés sur la base de test. En étudiant les annotations nous avons pu déterminer que cela était sûrement dû au nombre de locuteurs moyen par émission. Il est en effet plus important dans les émissions de la base de test que dans celles de développement (17,6 contre 15). De fait, plus de locuteurs secondaires n'étant pas correctement identifiés sur la base de test que sur la base de développement, les scores obtenus pour les métriques *semipond* et surtout *unipond* s'en trouvent fortement affectés.

émission	**NIST**	***unipond***	***semipond***
ensemble de développement			
CPB82055196	43.4 - 40.8	56.7	47.7
CPB84052346	35.4 - 32.0	39.0	33.7
CPB84052346 (part. 1)	30.2 - 27.1	37.9	31.8
CPB84052346 (part. 2)	40.0 - 36.4	36.3	34.7
CPB85104049	24.6 - 18.7	20.1	18.7
moyenne	34.5 - 30.5	38.6	33.4
ensemble de test			
CPB82051110	21.7 - 18.7	32.2	24.7
CPB82051645	35.2 - 29.8	64.7	50.6
CPB88000401	43.8 - 43.2	65.6	49.7
moyenne	33.6 - 30.6	54.2	41.7

Tableau 6.4 – Taux d'erreur pour la reconnaissance de locuteurs sur 6 émissions du *Grand Échiquier* pour les trois métriques (NIST, *unipond* et *semipond*). Les scores sont présentés sans prise en compte du phénomène de double-voix sauf pour la métrique NIST (avec - sans).

Afin d'écarter tout risque d'adaptation à un type particulier d'émissions de talk show, nous avons également testé l'algorithme *AV2* sur quatre émissions du corpus *On n'a pas tout dit*. Les tendances observées sur le *Grand Échiquier* semblent conservées comme en atteste le tableau 6.5. Le système utilisé est exactement identique à celui présenté à l'exception de la longueur minimale des plans lors de leur sélection (fixée à cinq secondes contre dix) et d'un changement du critère d'arrêt (seuil) pour le regroupement hiérarchique audio par distances probabilistes adapté sur une émission de développement. Le réglage de ces deux paramètres a été effectué en choisissant une des quatre émissions du corpus *On n'a pas tout dit* comme base de développement.

Il est intéressant de noter que les écarts de scores observés pour la métrique NIST avec et sans double-voix sont plus importants dans le tableau 6.5 que dans le tableau 6.4. Cela peut s'expliquer par le type de discours plus concis et succinct observé dans les émissions du corpus

On n'a pas tout dit pour lesquelles les intervenants se coupent la parole beaucoup plus fréquemment que dans les programme du *Grand Échiquier*.

émission	**NIST**	***unipond***	***semipond***
ensemble de développement			
OAPTD 1	35.8 - 27.7	30.8	29.5
ensemble de test			
OAPTD 2	38.2 - 25.6	32.9	30.4
OAPTD 3	49.0 - 42.5	64.5	55.5
OAPTD 4	27.9 - 25.4	39.0	35.0
moyenne	38.4 - 31.2	45.5	40.3

Tableau 6.5 – Taux d'erreur pour la reconnaissance de locuteurs sur 4 émissions d'*On n'a pas tout dit* pour les trois métriques (NIST, *unipond* et *semipond*). Les scores sont présentés sans prise en compte du phénomène de double-voix sauf pour la métrique NIST (avec - sans).

6.3 Exploitation de cohérence audiovisuelle : algorithme *AV3*

Un des défauts des algorithmes *AV1* et *AV2* est qu'aucun n'utilise d'information audiovisuelle pendant la phase de classification/resegmentation des données. Or, nous avons déjà indiqué à plusieurs reprises que les descripteurs vidéo sont notoirement plus stables temporellement que leurs homologues audio et peuvent donc fournir des décisions plus sûres. Eu égard aux comportements différents des caractéristiques issues des deux modalités, il est fort probable qu'une telle information pourrait être d'un grand intérêt. Nous proposons par conséquent d'implémenter un classifieur SVM pour l'information visuelle. Pour autant, cette information peut être délicate à manipuler dans le cas de la reconnaissance de locuteurs, une tâche qui n'implique a priori que la modalité audio. Par conséquent, afin de s'assurer de mesurer un accord entre les sorties de classification audio et vidéo, un calcul de cohérence est proposé comme cela est indiqué dans la figure 6.5.

étapes			
11	-	classification audio SVM (1 contre 1)	
12	-	évaluation des probabilités de sortie audio	
13	-	classification vidéo SVM (1 contre 1)	
14	-	évaluation des probabilités de sortie vidéo	*classification des trames de parole*
15	-	filtrage médian sur sorties audio et vidéo	
16	-	classification SVM des distributions jointes *AV* (1 classe)	
17	-	analyse des sorties de classifieurs et correction si lieu est	

Tableau 6.6 – Présentation des différentes étapes constitutives de la partie classification des trames de parole de l'algorithme *AV3*.

Repartant de l'algorithme *AV2* nous suggérons donc une version augmentée pour laquelle une approche de classification différente des trames de parole est proposée. Le tableau 6.6 en donne les étapes constitutives.

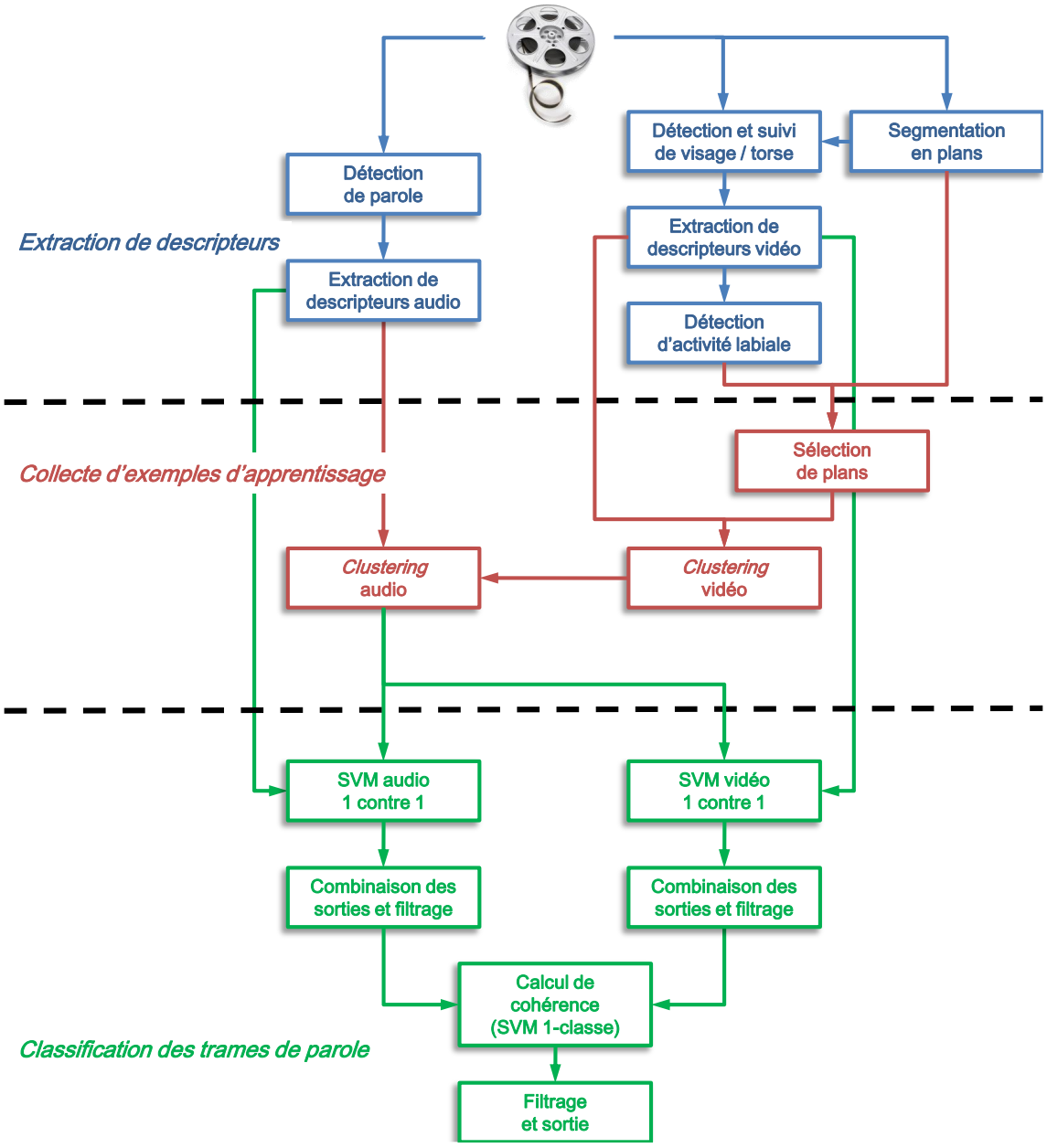


Figure 6.5 – Présentation de l'algorithme *AV3*.

6.3.1 Classification des trames de parole

La classification SVM des trames de parole est identique à celle présentée pour l'algorithme *AV2*. Cependant, nous proposons de joindre à celle-ci une classification vidéo basée sur les caractéristiques de costume.

6.3.1.1 Classification vidéo par SVM

De façon similaire à celle proposée pour la modalité audio, nous proposons d'effectuer une classification SVM sur les histogrammes de couleur de costume HSV extraits. Cependant, il est important de noter qu'ici, contrairement aux exemples récoltés lors de la constitution de la base d'apprentissage, toutes les trames de parole ne contiennent pas obligatoirement des informations de costumes. En effet, certains plans, tels les plans larges, ne présentent aucune détection de visage et de fait aucune région du torse n'a pu être extraite.

Par conséquent, des classifieurs SVM biclasses sont appris sur la base d'apprentissage obtenue précédemment et testés sur les seules trames de parole pour lesquelles une information de costume est disponible. Le processus de combinaison des sorties des $\frac{N(N-1)}{2}$ problèmes biclasses est le même que celui présenté plus haut. La seule différence est l'utilisation d'un autre noyau que celui présenté dans l'équation (6.4). En effet, à la place du traditionnel noyau gaussien, nous choisissons d'utiliser le noyau intersection d'histogramme, celui-ci étant particulièrement approprié à la comparaison d'histogrammes (comme son nom l'indique). La méthode d'intersection d'histogramme a été introduite par Swain et Ballard [1991] et le noyau résultant proposé entre autre par Maji *et al.* [2008]. Soit, avec Ha et Hb des histogrammes de b bins :

$$k(Ha, Hb) = \sum_{i=1}^{b} \min\{Ha_i, Hb_i\} \tag{6.6}$$

6.3.1.2 Indicateur de cohérence audiovisuelle

Comme nous l'avons indiqué plus haut, la modalité vidéo peut apporter de préciseuses informations sur l'action qui est montrée à l'écran et ainsi renforcer la prise de décision finale lors de l'attribution des étiquettes (*labels*) de locuteurs.

Méthodes existantes

L'idée d'étudier la corrélation audiovisuelle au cours du temps n'est pas nouvelle et de nombreuses méthodes ont été proposées. Ainsi, pour calculer la corrélation des sorties des classifieurs SVM audio et vidéo (sur les trames contenant bien évidemment les deux informations) nous avons envisagé d'utiliser des méthodes d'analyse de corrélation canonique (CCA) et d'analyse de co-inertie (CoIA). Ces dernières mesurent la relation linéaire entre deux variables multidimensionelles X_a et X_v (voir Hotelling [1936]). Elles permettent de trouver deux bases A et

B, une pour chaque variable, qui sont optimales au sens de la corrélation (respectivement de la covariance) ainsi que les corrélations (respectivement les covariances) maximales associées. Elles ont été utilisées avec succès dans des travaux de biométrie, en particulier ceux de Bredin et Chollet [2007] :

$$(A,B) = \text{argmax corr}(A^t X_a, B^t X_v) \qquad \text{CCA} \tag{6.7}$$

$$(A,B) = \text{argmax cov}(A^t X_a, B^t X_v) \qquad \text{CoIA} \tag{6.8}$$

De façon similaire, Li *et al.* [2003] proposent une analyse factorielle cross-modale (CFA). Les auteurs la présentent comme plus robuste qu'une analyse de corrélation canonique (CCA). Il s'agit ici de trouver les deux bases pour lesquelles :

$$(A,B) = \text{argmin } \|X_a A - X_b B\|_F^2 \qquad \text{CFA} \tag{6.9}$$

avec $A^t A = I$, $B^t B = I$ et $\|\cdot\|_F$ la norme de Frobenius. Dans tous les cas, les versions transformées de X_a et X_v sont ensuite obtenues suivant :

$$\begin{cases} \tilde{X}_a = X_a A \\ \tilde{X}_v = X_v A \end{cases} \tag{6.10}$$

Là encore, n'ayant pas eu à notre disposition d'annotation visuelle nous n'avons pu mesurer de façon effective les performances de détection de corrélation pour ces trois méthodes. Cependant, appliquées à notre base développement, ces méthodes se sont avérées très sensibles aux variations de couleur (et donc aux changements de plans) portées par les histogrammes HSV de costume. Ainsi, un fonctionnement en « registres » a été observé. En effet, les mesures de corrélation subissent un décalage (*offset*) différent pour chaque plan. Cela rend difficile le paramétrage d'un seuil au dessus duquel modalités audio et vidéo seraient corrélées et en dessous duquel elles ne le seraient pas.

Méta-descripteurs

Nous souhaitons construire, pour chaque trame de parole pour laquelle descripteurs audio et vidéo sont présents, un indicateur de confiance sur le choix de la classe proposée. Bien évidemment, en cas d'absence de descripteurs visuels de costume, la solution donnée par le classifieur audio est conservée. Nous nous plaçons ici dans le cas où chaque classifieur est en mesure de proposer une sortie de classification.

En étudiant pour chaque locuteur les distributions de probabilité de sortie des classifieurs SVM audio et vidéo (après filtrage), nous avons pu noter que celles-ci présentent des profils très

typiques. La figure 6.6 présente pour les données d'apprentissage d'une émission une visualisation « boîte à moustaches » (*box and whiskers plot*) de celles-ci. L'intérêt d'utiliser les distributions de sortie des classifieurs s'explique par le fait que les locuteurs principaux ont généralement tendance à bruiter les données des locuteurs secondaires (comme cela peut également être observé dans la figure 6.6). Cependant, des confusions propres à chaque locuteur peuvent être observées. Il est par conséquent possible d'apprendre pour chacun une distribution jointe audio-vidéo caractéristique qui peut être vue comme un méta-descripteur. En l'occurence, nous proposons de calculer ces distributions jointes sur les sorties reclassées des vecteurs de la base d'apprentissage. Ainsi, pour chacun des *clusters* nous créons un ensemble de méta-descripteurs associé.

La classification SVM à une classe

Nous proposons d'utiliser des classifieurs SVM à une classe pour déterminer si l'hypothèse fournie par l'un des classifieurs (à choisir) est correcte ou non. La classification SVM à une classe est une technique d'estimation de support de densité pouvant être considérée comme une alternative à l'estimation du minimum de volume (voir Schölkopf *et al.* [2001]). Cette dernière a pour but la minimisation du volume contenant une fraction d'exemples d'apprentissage tandis que la classification SVM à une classe utilise un critère de maximisation de la marge dans un espace de Hilbert à noyau reproduisant (RKHS).

Soit $\mathscr{X} = \{x_1, ..., x_l\}$ un ensemble de vecteurs d'apprentissage. L'astuce du noyau (*kernel trick*) permet de projeter les observations dans un espace de Hilbert $\mathscr{E}$ associé avec un noyau $k(\cdot,\cdot)$ et la transformation $\Phi(x) = k(x,\cdot)$ (RKHS). La technique de classification par SVM à une classe consiste à déterminer la fonction g_s décrivant un hyperplan $\mathscr{H}$ dans l'espace $\mathscr{E}$ et dont le signe est positif dans une région aussi petite que possible et qui capture la majorité des données. Cela est résolu en déterminant l'hyperplan

$$\mathscr{H} : \langle \omega, \Phi(x) \rangle - \rho \tag{6.11}$$

défini par le vecteur normal $\omega \in \mathscr{E}$ et le décallage ρ séparant les vecteurs de caractéristiques de l'origine avec une marge maximale et pour lequel $g_s(x) = \mathrm{sgn}\left(\langle \omega, \Phi(x) \rangle - \rho\right)$ avec sgn la fonction signe indiquant si un vecteur est du côté positif ou négatif de l'hyperplan. Cela revient à résoudre le problème suivant :

$$\min_{\omega,\xi,\rho} \quad \frac{1}{2}\|\omega\|^2 + \frac{1}{\nu l}\sum_i \xi_i - \rho \tag{6.12}$$

sous contrainte

$$\langle \omega, \Phi(x_i) \rangle \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \rho \geq 0 \tag{6.13}$$

avec ξ_i les variables introduites pour mesurer le degré d'erreur de classification, $1 \leq i \leq l$ et ν un paramètre de pénalisation positif permettant un compromis entre maximisation de marge et erreurs d'apprentissage.

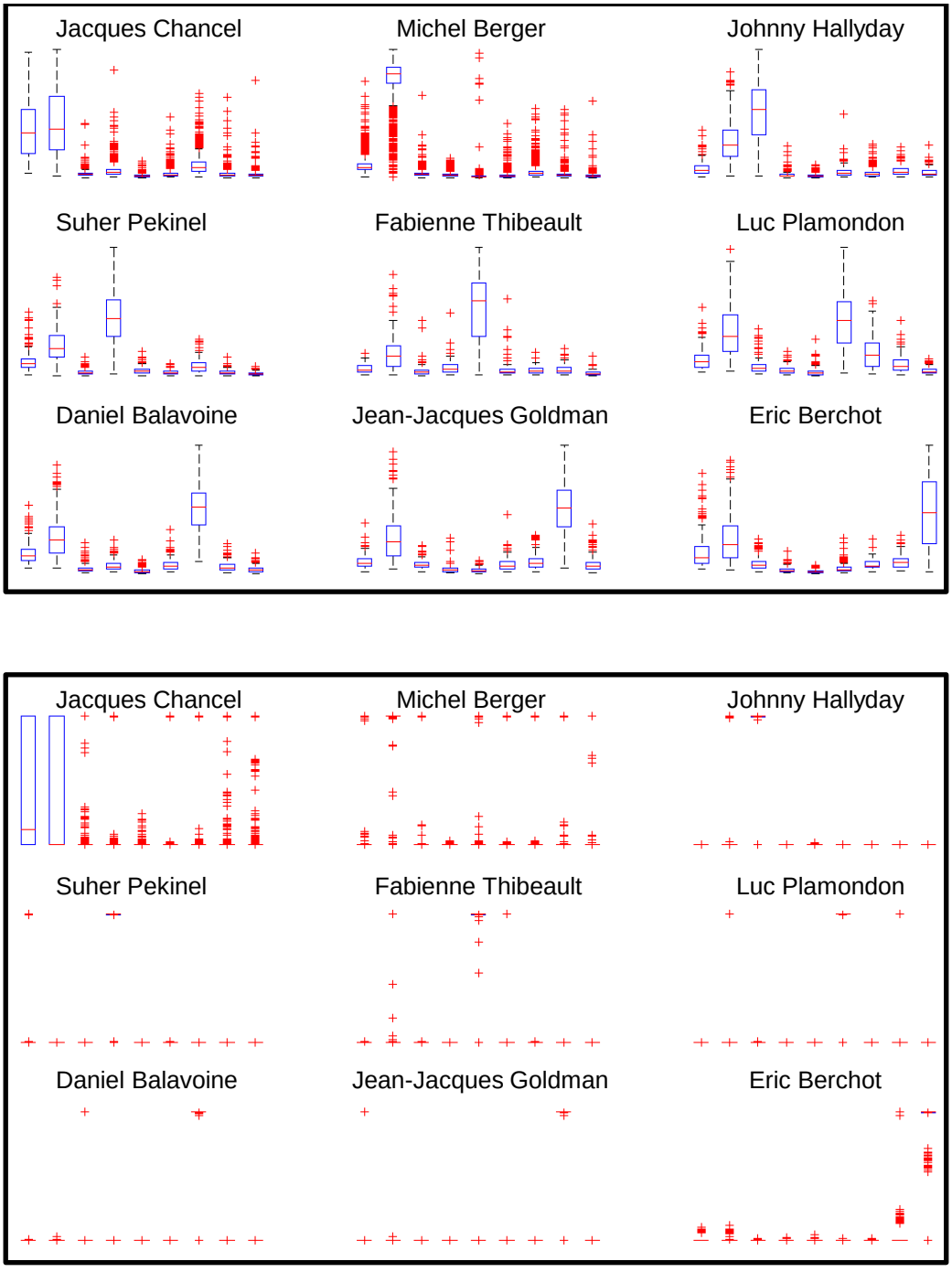


Figure 6.6 – Visualisation des distributions des sorties probabilisées des classifieurs audio (en haut) et vidéo (en bas) pour chaque locuteur sur les données d'apprentissage de l'émission CPB85104049 du corpus *Le Grand Échiquier*. Les colonnes représentent les modèles (*clusters*) obtenus à l'issue de la phase de collecte d'exemples d'apprentissage.

La solution est donnée par :

$$g_s(x) = \text{sgn}\left(\sum_{i=1}^{l} \alpha_i k(x_i, x) - \rho\right) \tag{6.14}$$

les α_i étant les multiplicateurs de Lagrange vérifiant $0 \leq \alpha_i \leq \frac{1}{\nu l}$. Les vecteurs de caractéristiques x_i pour lesquels $0 \leq \alpha_i \leq \frac{1}{\nu l}$ sont les vecteurs de support et ceux pour lesquels $\alpha_i = \frac{1}{\nu l}$ sont les données abérrantes (*outliers*). Les vecteurs de $\mathscr{X}$ restants ont des multiplicateurs de Lagrange de valeur nulle (voir figure 6.7).

Le paramètre ν joue un rôle central. Il est prouvé dans Schölkopf *et al.* [2001] que celui-ci est une borne supérieure de la proportion d'erreurs de classification et une borne inférieure de la proportion de vecteurs de support.

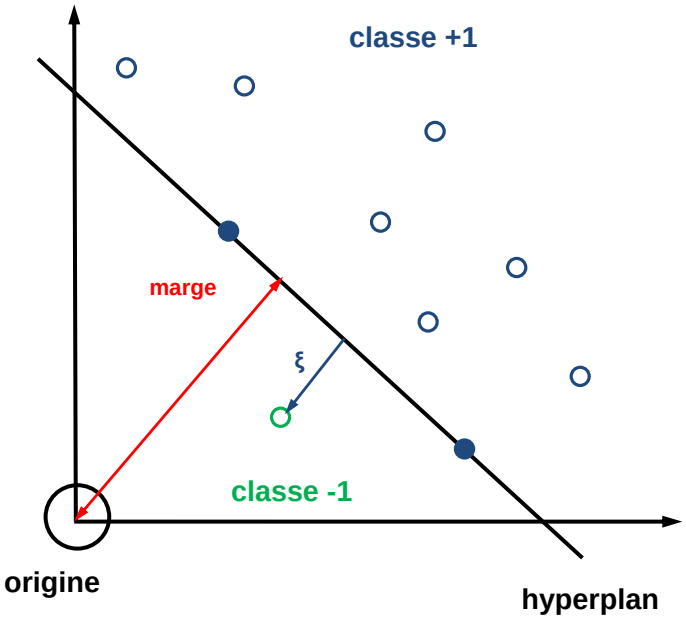


Figure 6.7 – Classification SVM à une classe avec maximisation de la marge séparatrice. Les vecteurs de support sont les points sur lesquels s'appuie la marge.

Enfin, il est intéressant de mentionner que l'optimisation présentée peut être résolue efficacement en utilisant une variante des méthodes de sélection de sous-ensembles (*subset method selection*) comme proposé dans Schölkopf et Smola [2001]. Dans notre travail nous avons utilisé l'implémentation disponible dans la boîte à outils *LIBSVM*.

Indicateur de cohérence audiovisuelle

Dans notre étude, sont appris n classifieurs SVM à une classe (un par *cluster* d'apprentissage obtenu). Les données fournies sont pour chaque trame les méta-descripteurs présentés plus haut, soit la concaténation des sorties probabilisées des classifieurs audio et vidéo. Nous choisissons l'un des deux classifieurs pour nous fournir, pour chaque trame de parole, une étiquette hypothétique (toujours de façon non-supervisée). Les *outliers* détectés sont alors supposés correspondre aux situations de décalage entre modalité audio et vidéo, c'est-à-dire lorsque la personne à l'écran ne correspond pas au locuteur actif ou vice versa.

En raison de sa stabilité au cours du temps, notre choix s'est porté sur le classifieur vidéo. Ensuite, nous classifions le méta-descripteur de chaque trame de parole considérée par le classifieur SVM correspondant à l'étiquette hypothétique. Si la valeur est supérieure à un seuil (théoriquement zéro), alors le choix de l'étiquette donnée par le classifieur vidéo est entériné puisque le méta-descripteur de la trame considérée appartient à la distribution. Si elle est inférieure, il y a désaccord. Nous choisissons alors de prendre l'étiquette de sortie du classifieur audio (le plus sûr puisque la tâche considérée est audio). Nous avons laissé le seuil de l'indicateur à zéro et le paramètre ν a été fixé à 0.2. La figure 6.8 résume la méthode proposée.

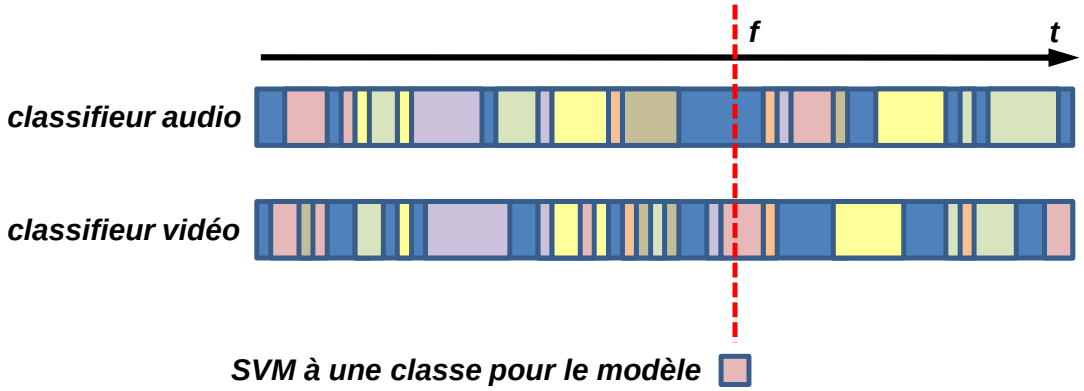


Figure 6.8 – Schéma illustratif de l'indicateur de cohérence audiovisuelle pour une trame f. Le SVM à une classe calculé pour le modèle mauve est appliqué pour déterminer si la trame f est correctement classée ou pas.

La figure 6.9 propose une illustration des résultats obtenus avec les méthodes de détection de cohérence évoquées. Sont ainsi indiquées, les méthodes état de l'art d'analyse de corrélation canonique (CCA), d'analyse de co-inertie (CoIA) et d'analyse factorielle cross-modale (CFA). De plus, est tracé l'indicateur présenté plus haut (SVM à une classe). Les changements de plans apparaissent en traits pleins et noirs ; les changement de locuteurs en traits fins et pointillés. Plus bas, une trame correspondant au plan courant est étirée suivant la durée de ce plan et en dessous sont indiqués les tours de parole (ici avec deux locuteurs : Jacques Chancel et Johnny Hallyday). En blanc sur cette partie de la figure apparaît le phénomène de double-voix.

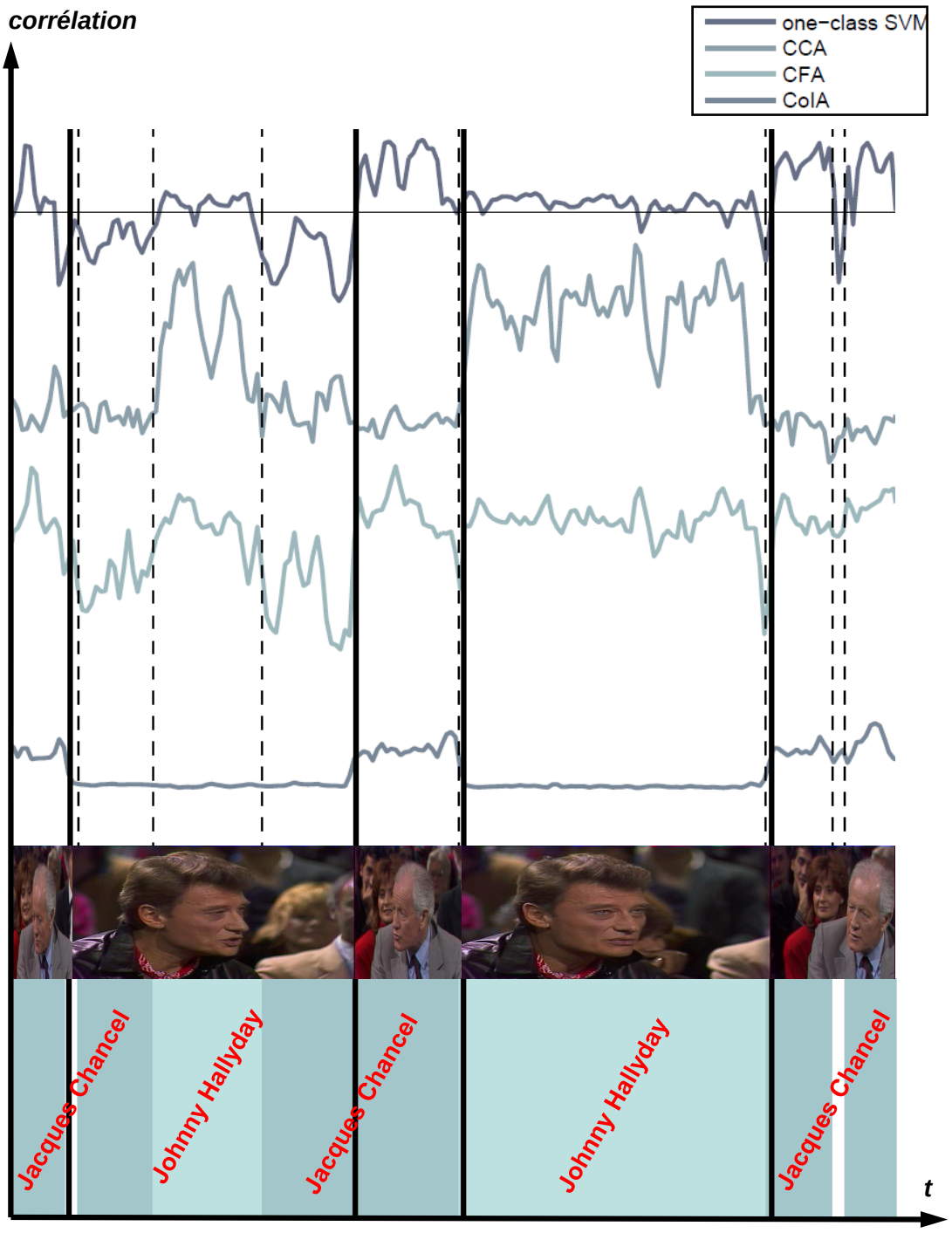


Figure 6.9 – Comparaison de méthodes mesurant la corrélation audiovisuelle (SVM à une classe, CCA, CoIA et CFA) pour une partie de dialogue de l'émissions CPB85104049 du corpus *Le Grand Échiquier*.

La figure illustre que les méthodes CFA et SVM à une classe donnent les résultats les plus pertinents alors qu'un fonctionnement en « registre » est observé pour CCA et surtout CoIA. Cependant, ces variations observées lors des changements de plans avec les méthodes CFA (peu visibles ici) sont sensiblement plus marquées et rendent l'élaboration d'un seuillage beaucoup plus complexe que pour le classifieur SVM à une classe.

6.4 Résultats et discussion

Comme en témoigne le tableau 6.7, le système *AV3* montre une amélioration significative des scores de reconnaissance de locuteurs, que l'évaluation soit effectuée avec les métriques NIST, *unipond* ou *semipond*.

émission	**NIST**	***unipond***	***semipond***
ensemble de développement			
CPB82055196	40.9 - 37.9	57.5	47.7
CPB84052346	32.9 - 29.3	38.3	32.1
CPB84052346 (part. 1)	27.7 - 24.4	37.1	30.6
CPB84052346 (part. 2)	37.5 - 33.8	37.4	33.9
CPB85104049	21.8 - 15.5	19.9	17.9
moyenne	31.9 - 27.6	38.5	32.6
ensemble de test			
CPB82051110	22.1 - 18.7	32.0	24.2
CPB82051645	31.9 - 26.1	64.5	49.7
CPB88000401	44.4 - 43.4	63.9	48.6
moyenne	32.8 - 29.4	53.5	40.8

Tableau 6.7 – Taux d'erreur pour la reconnaissance de locuteurs sur 6 émissions du *Grand Échiquier* pour les trois métriques (NIST, *unipond* et *semipond*). Les scores sont présentés sans prise en compte du phénomène de double-voix sauf pour la métrique NIST (avec - sans).

La métrique *unipond* est celle pour laquelle l'amélioration est la moins notable, ce qui pouvait être attendu puisque les poids des différents locuteurs sont équilibrés. En effet, l'initialisation des *clusters* étant inchangée, aucun nouveau modèle de locuteur n'apparaît. L'amélioration apportée par la correction audiovisuelle est alors assez minime. Il est de plus notable que l'approche de correction audiovisuelle par calcul de cohérence améliore les résultats pour toutes les émissions à l'exception de CPB88000401 (avec la métrique NIST) pour laquelle ils sont légèrement déteriorés. Néanmoins, cette dégradation est très marginale et le score moyen de l'ensemble de test ne s'en trouve pas fortement altéré.

La figure 6.10 montre une matrice de confusion des trames de parole après classification. Sur cette émission, on peut observer un nombre équivalent de locuteurs et de *clusters*. De plus, chaque *cluster* représente de façon effective un locuteur donné. Ceci est rendu visible par l'observation de la diagonale centrale bien marquée. Les confusions les plus fréquentes concernent les *clusters* associés aux deux locuteurs principaux (Jacques Chancel et Michel Berger). Celles-ci étant les plus importantes au sens du nombre d'éléments, elles ont, par phénomène de gravité,

plus tendance à attirer à elles les trames correspondants à d'autres locuteurs. En effet, plus un *cluster* est de taille importante et plus il aura tendance à contenir des éléments de nature variée et ainsi à en attirer d'autres à lui. De plus le présentateur et l'invité principal étant les points de focalisation de l'émission, il arrive bien souvent qu'ils soient montrés à l'écran alors qu'un autre intervenant prend la parole, ce qui peut entraîner des erreurs. Il est également intéressant d'observer ce schéma pour se faire une idée plus précise de la répartition de la parole dans une émission de talk show. En effet, on remarque très nettement les deux locuteurs principaux et on peut noter les faibles interventions de certains locuteurs secondaires (Patrice Vigier et Fabienne Thibeault par exemple).

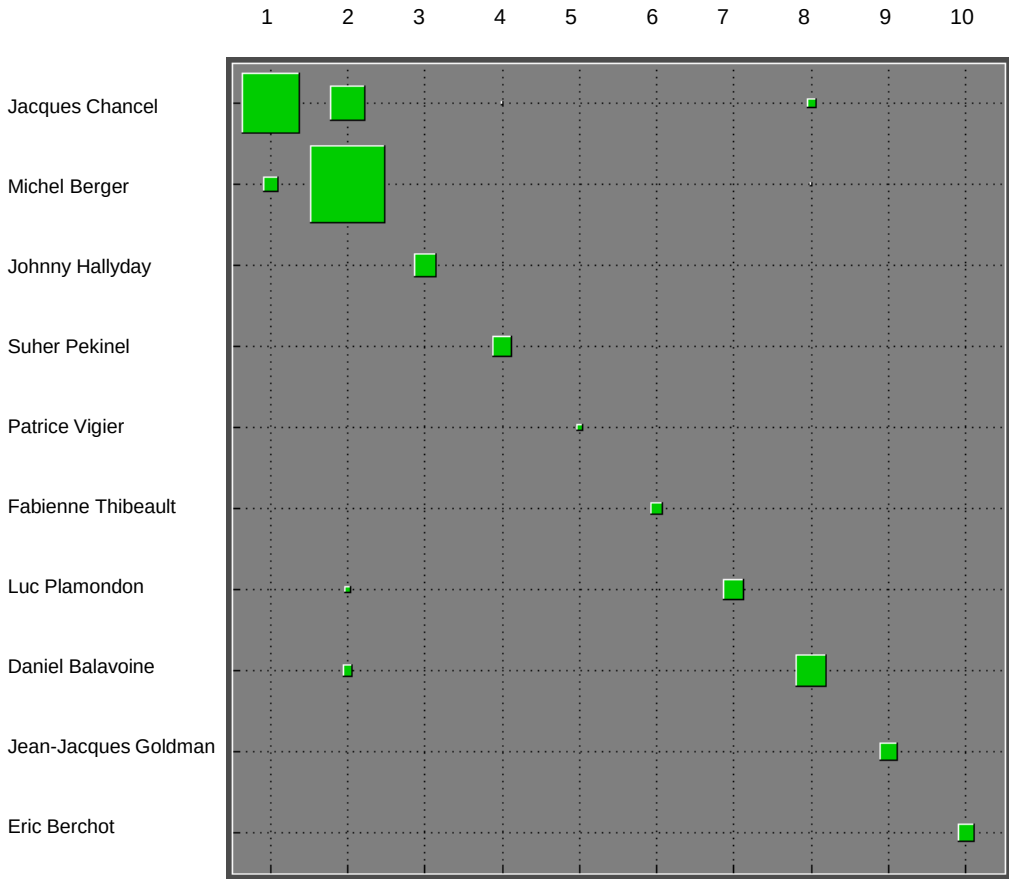


Figure 6.10 – Matrice de confusion indiquant les associations locuteurs/*clusters* (lignes/colonnes) réalisées par le système de classification pour l'émission CPB85104049 du corpus *Le Grand Échiquier*. La taille des rectangles verts est proportionnelle au nombre de trames appartenant à chaque *cluster*.

Une autre matrice de confusion est présentée à la figure 6.11 pour l'émission CPB84052346. Cette émission comporte un total de quinze locuteurs et la répartition des trames de parole

semble cohérente avec les scores du tableau 6.7. Là encore, un nombre équivalent de *clusters* et de locuteurs a été trouvé. Cependant, contrairement à la figure 6.10, certains *clusters* représentent le même locuteur. C'est particulièrement flagrant pour François Périer dont les trames de parole sont principalement réparties dans les *clusters* 5 et 11. On peut également noter que, là encore, certains locuteurs parlent extrêmement peu : Stéphane Grappelli, François-René Duchâble, Farid Chopel, etc., et que l'étalement des *clusters* représentant les locuteurs majoritaires (Jacques Chancel et Gérard Oury) est plus important que pour l'émission précédente, ceci s'expliquant certainement par la moins bonne définition des modèles de locuteurs à l'issue de la phase d'apprentissage non-supervisée.

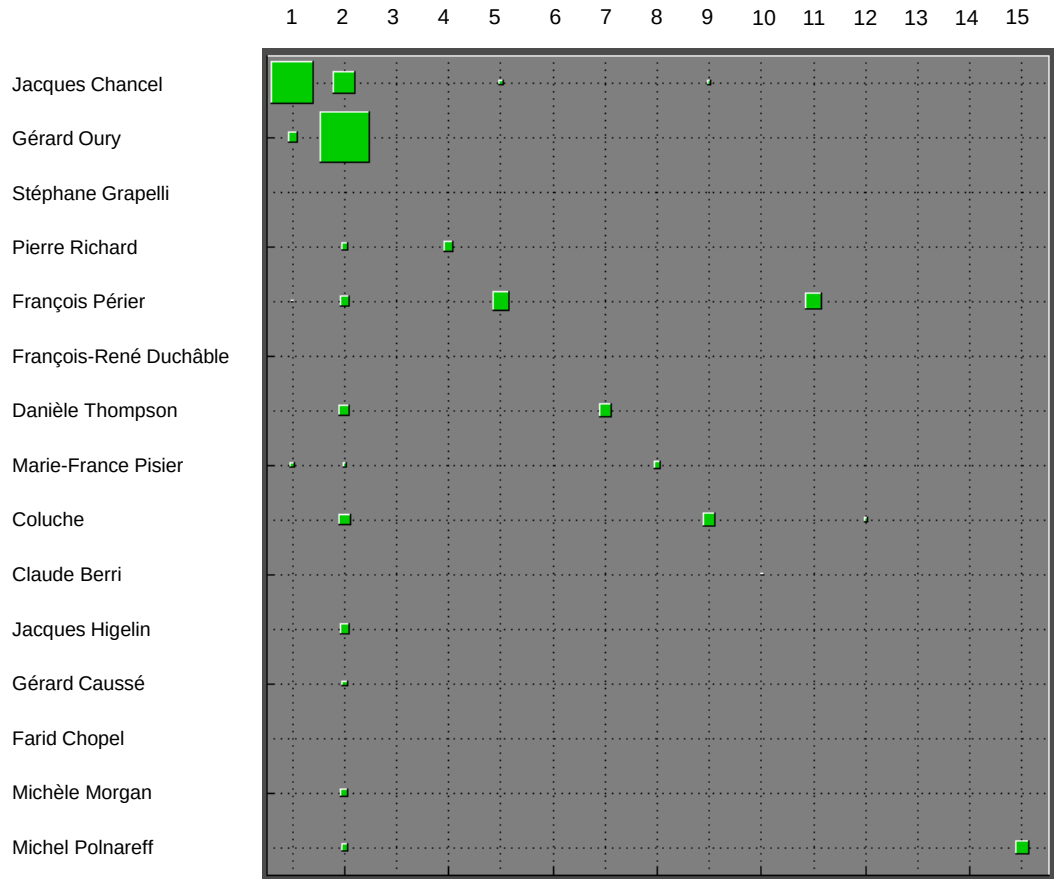


Figure 6.11 – Matrice de confusion indiquant les associations locuteurs/*clusters* (lignes/colonnes) réalisées par le système de classification pour l'émission CPB84052346 du corpus *Le Grand Échiquier*. La taille des rectangles verts est proportionnelle au nombre de trames appartenant à chaque *cluster*.

Dans le dernier exemple de matrice de confusion présenté figure 6.12 on peut noter que les aspects évoqués précédemment sont également présents. Il est intéressant d'observer que dans

ce cas, pour les vingts locuteurs intervenant au cours de l'émission, seules seize *clusters* d'apprentissage ont été proposés. Par conséquent, les métriques *unipond* et *semipond* s'en trouvent affectées avant même que ne commence la phase de classification. En particulier, pour *unipond* le taux d'erreur de reconnaissance ne pourra pas se situer en dessous de 20% ($\frac{20-16}{20}$).

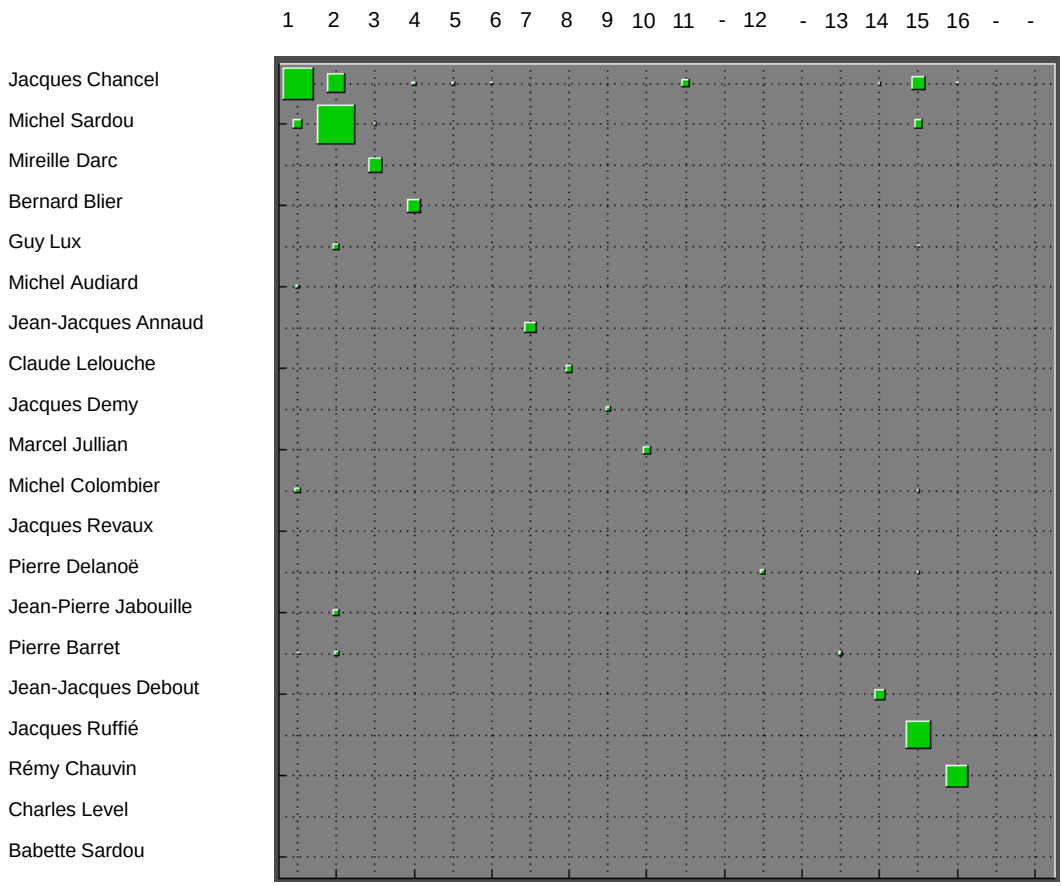


Figure 6.12 – Matrice de confusion indiquant les associations locuteurs/*clusters* (lignes/colonnes) réalisées par le système de classification pour l'émission CPB82055196 du corpus *Le Grand Échiquier*. La taille des rectangles verts est proportionnelle au nombre de trames appartenant à chaque *cluster*.

De plus, bon nombre de ces *clusters* représentent le même locuteur (les *clusters* 1, 5 et 11 pour Jacques Chancel par exemple). L'étalement des *clusters* majoritaires semble moins important que pour l'émission CPB84052346. Les faibles résultats de reconnaissance sur cette émission (en particulier pour les métriques *unipond* et *semipond*) semblent plutôt dus au grand nombre de locuteurs qui ne sont pas correctement reconnus (Jacques Revaux, Babette Sardou, Charles Level, etc.). Enfin, il est intéressant de remarquer la confusion présente au *cluster* 15 entre Jacques Chancel et Jacques Ruffié. Celle-ci est assez importante et peut être expliquée

entre autres par la couleur de costume très proche des deux intervenants comme en témoigne la figure 6.13. Par conséquent, il serait d'intérêt de fournir des descripteurs non colorimétriques permettant leur différenciation (ceux de mouvement par exemple).

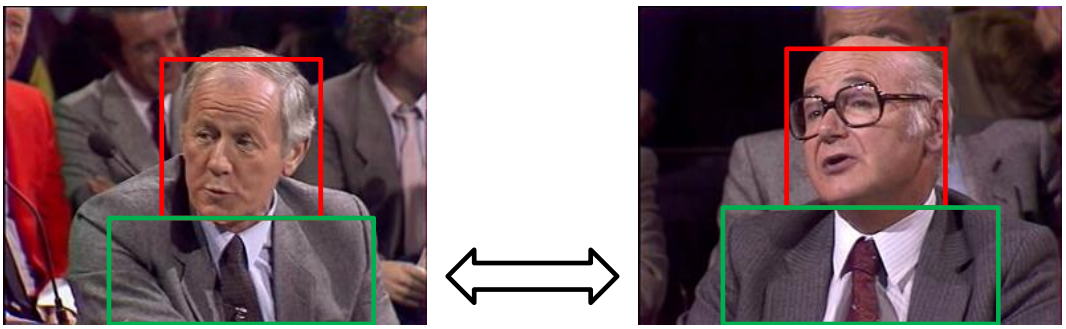


Figure 6.13 – Comparaison de la couleur du costume pour Jacques Chancel et Jacques Ruffié pour l'émission CPB82055196 du corpus *Le Grand Échiquier*.

Comme précédemment, afin de valider l'algorithme *AV3*, nous proposons d'effectuer la tâche de reconnaissance de locuteurs sur quatre émissions du corpus *On n'a pas tout dit*. Les taux d'erreur NIST, *unipond* et *semipond* sont donc calculés. Les résultats obtenus sont ici aussi cohérent avec la tendance dégagée de ceux du *Grand Échiquier*, avec une amélioration significative des résultats suite à l'ajout du module de calcul de cohérence audiovisuelle. De plus, comme pour les résultats obtenus avec le système *AV2*, on peut observer une différence notable entre les scores NIST avec et sans double-voix, ces émissions contenant de nombreuses interventions de très courte durée ainsi que beaucoup de segments de double-voix.

émission	**NIST**	***unipond***	***semipond***
ensemble de développement			
OAPTD 1	33.3 - 24.7	27.6	27.1
ensemble de test			
OAPTD 2	37.6 - 21.0	33.6	26.1
OAPTD 3	48.9 - 42.3	63.1	55.3
OAPTD 4	26.9 - 24.2	37.3	32.8
moyenne	37.8 - 29.2	44.7	38.0

Tableau 6.8 – Taux d'erreur pour la reconnaissance de locuteurs sur 4 émissions d'*On n'a pas tout dit* pour les trois métriques (NIST, *unipond* et *semipond*). Les scores sont présentés sans prise en compte du phénomène de double-voix sauf pour la métrique NIST (avec - sans).

Conclusion

Nous avons proposé dans ce chapitre un système de reconnaissance de locuteurs complètement nouveau. Celui-ci se mesure avantageusement avec les résultats obtenus avec les algorithmes état de l'art (ou légèrement dérivés) présentés au chapitre 5. Nous en avons proposé deux versions (*AV2* et *AV3*) aux fonctionnements légèrement différents, le second étant une extension du premier. La partie commune des deux systèmes consiste à sélectionner très méticuleusement des exemples pour constituer une base d'apprentissage non-supervisée de la meilleure qualité possible. Pour ce faire, seuls sont conservés les plans suffisamment longs et pour lesquels un locuteur actif apparaît à l'écran. Des regroupements hiérarchiques successifs vidéo et audio permettent ensuite la création de modèles de locuteurs très fiables. Une classification SVM est alors effectuée pour étiqueter toutes les trames de parole. La version *AV3* consiste au simple ajout d'un classifieur SVM sur les données vidéo et d'un module de correction par mesure de cohérence audiovisuelle. Ce dernier constituant permet d'évaluer si les sorties probabilisées fournies par les classifieurs audio et vidéo sont en accord. Dans le cas d'une réponse positive, la sortie plus stable du classifieur vidéo est gardée. Dans celui d'une réponse négative, le choix de l'attribution de l'étiquette de sortie est donné par le classifieur audio seul.

L'amélioration successive des résultats avec les trois métriques proposées au chapitre 4 pour les différents systèmes a conforté l'idée que l'utilisation de caractéristiques visuelles pouvait être d'un grand intérêt, même pour la résolution de tâches supposémment exclusivement audio comme la reconnaissance de locuteurs. De plus, nous avons montré que de très bons résultats peuvent être obtenus avec des méthodes à noyaux pour la réalisation d'un système de reconnaissance de locuteurs.

AYANT proposé une approche de structuration d'émissions de talk show non-supervisée et basée sur la connaissance des tours de parole des locuteurs dans la première partie, nous avons apporté une attention particulière dans la suite de cette thèse à la tâche de reconnaissance de locuteurs (*speaker diarization*). Ainsi, après avoir proposé un tour d'horizon des méthodes existantes, nous avons étudié les résultats obtenus avec un système état de l'art sur le type de contenu particulier que sont les émissions de talk show. Ces résultats étant bien moins convainquants que ceux mesurés sur les corpus de référence pour la reconnaissance de locuteurs, nous nous nous sommes par la suite concentrés à proposer de nouvelles approches pour cette tâche.

Dans une première étude nous avons suggéré l'idée de joindre une méthode d'initialisation audiovisuelle des modèles de locuteurs à un système de reconnaissance état de l'art. Partant de l'hypothèse que le locuteur actif est dans la plupart des cas montré au téléspectateur dans une émission de talk show, cette initialisation a été réalisée en extrayant des informations visuelles des individus à l'écran. En particulier, des caractéristiques colorimétriques des costumes des intervenants ont été calculées. L'étude des résultats obtenus sur le corpus *Le Grand Échiquier* a montré une amélioration de la reconnaissance des locuteurs par rapport au sytème de base.

Enfin, nous avons proposé un système de reconnaissance original basé sur l'utilisation de méthodes à noyau. Une base d'apprentissage est ainsi constituée de façon non-supervisée. Tout d'abord, au moyen d'un détecteur d'activité labiale, en sélectionnant les plans pour lequels un locuteur actif est à l'écran. Puis en effectuant des regroupements hiérarchiques vidéo et audio de ces plans. Un classifieur SVM audio appris sur les modèles de locuteurs ainsi constitués est ensuite utilisé (système *AV2*). Dans une variante de ce système nous proposons de joindre un classifieur SVM vidéo et de calculer un indicateur de cohérence audiovisuelle par classifieur SVM à une classe. Suivant le score de celui-ci, la sortie du SVM audio ou vidéo est choisie. La validité du système est ensuite confortée par les résultats obtenus sur les corpus *Le Grand Échiquier* et *On n'a pas tout dit.*

Conclusions et perspectives

Conclusions

Le travail mené au cours de cette thèse a été l'occasion de proposer de nouvelles techniques pour l'indexation, la segmentation et la structuration automatique de documents audiovisuels. En particulier, nous avons pu mesurer l'apport de l'utilisation de modalités multiples, en l'occurence audio et vidéo, à une tâche de structuration. De plus, ayant été confrontés directement aux problèmes rencontrés par une entreprise d'archivage de contenus audiovisuels (Ina), nous avons été en mesure d'apprécier les attentes des utilisateurs concernant les systèmes d'indexation automatiques.

Cette étude a ainsi permis l'obtention d'un schéma de structuration non-supervisé et générique des émissions de talk show. Pour cela, nous avons dans un premier temps étudié la nature même de ce genre télévisuel avant d'en proposer une structuration. Les composantes caractéristiques de ce genre d'émission ont été mises en évidence par l'étude de travaux sémiologiques. Nous avons ainsi noté l'importance de la parole véhiculée par le couple présentateur(s)/invité(s), pivot de ce type de programme. De plus, nous avons observé les codes très spécifiques qui sont employés, en particulier par le présentateur dont le rôle peut être défini comme celui d'un chef d'orchestre. Une étude comparative des deux talk shows *Le Grand Échiquier* et *On n'a pas tout dit* nous a ensuite permis de constater que malgré les différences observées ces traits caractéristiques sont effectivement partagés.

Fort de ce constat, nous avons proposé une organisation générique des émissions de talk show. Pour cela nous avons mis en avant l'idée que l'élaboration d'un schéma de structuration est nécessairement tributaire d'une application ou cas d'usage. En ayant suggéré quelques-uns (comme la ré-éditorialisation, la création automatique d'une table des matières ou l'élaboration de logiciels d'annotation semi-automatisés), nous avons proposé de regrouper les invariants des programmes de talk show en trois grandes familles : « contenu », « délimiteurs » et « localisation ». Nous avons ensuite constaté la validité d'un tel schéma au moyen d'une expérience utilisateur pour laquelle la structuration suggérée permet une navigation plus rapide.

Ayant ainsi identifiés les éléments constitutifs des émissions de talk show, nous avons ensuite proposé un tour d'horizon des méthodes état de l'art utiles à leurs détection automatisée. Pour cela a été effectuée une distinction entre méthodes de segmentation (segmentation en plans, segmentation audio, etc.), de détection de « concepts » de haut-niveau (détection d'applaudissements, détection de rires, etc.) et détection de « concepts » de niveau supérieur (détection d'inserts, détection de performances non-musicales parlantes, etc.). Ces considérations ont mis en avant l'importance de l'identification des tours de parole. De fait, nous avons proposé de placer ceux-ci au centre de notre architecture de structuration et de nous concentrer sur les méthodes de reconnaissance de locuteurs (*speaker diarization*).

Après avoir donné un aperçu des techniques de reconnaissance de locuteurs, nous avons testé la validité d'un système état de l'art pour la reconnaissance automatique et non-supervisée

des intervenants d'émissions de talk show. À la vue des résultats, nous avons remarqué que ce type de systèmes, bien souvent paramètré pour être utilisé lors de campagnes d'évaluation (de type NIST), ainsi que les méthodes d'évaluation, sont relativement mal ajustés à la tâche que nous souhaitons effectuer. Par conséquent, nous avons suggéré de nous concentrer sur l'élaboration de systèmes mieux appropriés à la reconnaissance d'intervenants de talk show et proposé l'utilisation de deux métriques d'évaluation des résultats.

Dans un premier temps, nous avons soumis l'idée d'adapter un système état de l'art à la tâche de reconnaissance en introduisant une étape d'initialisation fondée sur l'information visuelle véhiculée par les costumes portés par les intervenants. Les résultats obtenus ont confirmé l'intérêt d'utiliser des descripteurs visuels pour ce type de contenu pour lequel la modalité vidéo peut apporter un complément d'information utile à la modalité vidéo.

Puis, dans un second temps nous avons proposé l'élaboration d'un système de reconnaissance original. Celui-ci se base sur l'utilisation de méthodes à noyau, habituellement peu employées pour la tâche de reconnaissance de locuteurs. A partir d'une sélection de plans pour lesquels la personne à l'écran est supposée être la personne qui parle, nous avons suggéré d'utiliser une succession de deux regroupements hiérarchiques. Le premier est un *clustering k*-moyennes sous-optimal pour lequel les plans de contenus visuels similaires sont aggrégés alors que le second exploite les caractéristiques audio des partitions nouvellement créées pour effectuer un nouveau regroupement par calcul de distances probabilistes en RKHS (espace de Hilbert à noyau reproduisant). Puis, ayant ainsi constitué de manière non-supervisée des données d'apprentissage fiables des locuteurs présents sur le plateau de l'émission, nous avons procédé à une classification audio par machines à vecteurs de support (SVM). Nous avons ensuite proposé l'adjonction à cette architecture d'un module de correction d'erreur de classification par calcul de cohérence entre les sorties SVM audio et vidéo. Cette mesure, effectuée à l'aide de classifieurs SVM à une classe, a permis de renforcer la prise de décision en indiquant pour chaque trame de parole laquelle des deux propositions était à conserver. Comparativement à l'algorithme état de l'art et à sa version augmentée précédemment présentés, ce système a obtenu de très bons résultats sur le corpus *Le Grand Échiquier*. Ces résultats ont ensuite été confirmés sur un autre corpus de test assez different (*On n'a pas tout dit*).

Perspectives

Dans Vallet *et al.* [2010] nous avons observé l'avantage d'utiliser des descripteurs de mouvement pour caractériser les intervenants. En effet, ceux-ci peuvent s'avérer particulièrement précieux pour désambiguiser des confusions résultantes de l'utilisation unique de la couleur du costume (dans le cas où deux locuteurs sont habillés de façon similaire). En effet, les individus prenant la parole peuvent montrer des attitudes très différentes : agité, calme, avec beaucoup de mouvements de mains, etc. Une implémentation de ces caractéristiques dans le système final semble par conséquent très prometteuse.

Dans la même étude exploratoire nous avons montré que des performances intéressantes peuvent être attendues d'un système semi-automatique, c'est à dire pour lequel des exemples

d'apprentissage pour chaque locuteur sont fournis en amont par l'utilisateur (un documentaliste de l'Ina par exemple). Un tel dispositif présente de plus l'avantage d'éviter l'étape d'association d'un locuteur à chaque *cluster* qui peut parfois être délicate. En effet, en plus du coup imposé à l'utilisateur, un inconvenient majeur des systèmes de reconnaissance de locuteurs complètement non-supervisés est que certains locuteurs puissent être introuvables, répartis sur plusieurs *clusters*, etc. les systèmes n'étant pas parfaits. Par conséquent, une approche semi-automatique ne semble dès lors guère plus coûteuse en termes d'efforts pour l'utilisateur.

Il serait également envisageable de proposer à l'utilisateur de valider les bases d'apprentissage constituées automatiquement dans le dernier système présenté. Ainsi, à l'aide de mesures de confiance, le système pourrait proposer de vérifier si l'identité du locuteur correspond bien à celle attendue dans quelques séquences problématiques (apprentissage actif). Les modèles ainsi constitués seraient alors d'autant plus à même de fournir des données fiables pour l'apprentissage de classifieurs discriminants. Qui plus est, l'utilisateur serait en mesure de fournir des exemples d'apprentissage dans le cas où aucun modèle n'est créé pour un intervenant donné.

Le traitement du phénomène de double-voix dans les travaux de reconnaissance de locuteurs est un problème important puisqu'il n'existe à ce jour aucune méthode convainquante pour sa détection. Lors de l'introduction de l'indicateur de cohérence entre classifieurs audio et vidéo, nous avons observé que la plupart du temps ces segments de parole étaient bien détectés (c'est-à-dire avec une mesure largement en dessous du seuil fixé). Par conséquent, un travail approfondi pourrait permettre la détection de ce phénomène dans le cas de contenu vidéo.

Dans sa thèse récemment soutenue (voir Bendris [2011]), Meriem Bendris propose une indexation audiovisuelle des intervenants pour l'émission de talk show *On n'a pas tout dit*. La comparaison de ce processus de sélection avec celui proposé dans notre dernier système pourrait être d'intérêt pour améliorer encore la constitution non-supervisée de base d'apprentissage.

De plus, il serait bien évidemment possible d'étudier les avantages d'avoir recours à des méthodes de traitement du langage pour proposer de façon automatique des labels d'intervenants, d'œuvres, etc. Pour cela il pourrait être utile d'avoir recours à la transcription automatique de parole ou à la reconnaissance automatique de caractères (OCR, *Optical Character Recognition*) ou encore les technologies de traitement automatique du langage naturel (TALN) pour l'analyse des notices descriptives. Une étude préliminaire a d'ailleurs été effectuée dans ce sens en utilisant les transcriptions de parole fournies par VECSYS et les notices documentaires dans le cadre du projet Infom@gic pour étiquetter automatiquement les auteurs de *performances musicales* (voir Vallet *et al.* [2008]).

Enfin, dans ses travaux de thèse (voir par exemple Bigot *et al.* [2010a] et Bigot *et al.* [2010b]), Benjamin Bigot propose d'effectuer la structuration et l'indexation de documents audiovisuels en déterminant le rôle des locuteurs (présentateur, invité, chroniqueur, journaliste, etc.). Pour cela, des descripteurs audio (acoustiques, prosodiques et temporels) sont utilisés. La connaissance de ces différents rôles pourrait permettre la détection de « concepts » de niveau supérieur du type de ceux que nous avons présenté de façon exploratoire avec la détection de *performances non-musicales parlantes* (comme les scènes de théatre, les tirades, les sketches, etc.).

Annexes

Annexe A

Description technique des corpus

Corpus *Le Grand Échiquier*

Pour constituer ce corpus, les émissions sélectionnées ont du être montées à la main. En effet, l'Ina stocke les émissions sous la forme de segments d'une heure légèrement recouvrant. Il a donc été nécessaire de reconstituer les émissions au moyen d'un logiciel de montage. Dans ce cas, le logiciel *MPEG2 Editor*, permettant de coller et découper des parties de vidéos sans nécessiter leur encodage, a été utilisé. Deux formats de compression sont utilisés par l'Ina pour le traitement des archives numérisées : MPEG-1 et MPEG-2. Les archives sont toujours sauvées dans les deux formats, la version MPEG-2 étant la version de haute qualité alors que la version MPEG-1 est une version dégradée et plus légère utilisée généralement pour la visualisation de contenus (en particulier sur internet). Le corpus a été constitué au moyen de vidéos MPEG-1. Plus précisément, les propriétés des vidéos ainsi créées sont données dans le tableau A.1 :

VIDEO	
Dimensions	352 × 288 pixels
Codec	MPEG-1 video
Fréquence d'aquisition	25 images/seconde
Débit	~ 1100 kb/s
AUDIO	
Codec	MPEG-1 Audio, couche 2
Canaux	stéréo
Fréquence d'échantillonnage	48000 Hz
Débit	96 kb/s

Tableau A.1 – Propriétés des émissions du corpus *Le Grand Échiquier*.

Un des désavantages du corpus ainsi constitué est qu'outre sa qualité moyenne, la piste audio est sur-échantillonnée suite à des pré-traitements antérieurs. Par conséquent, la piste audio a été extraite et basculée de stéréo à mono avant de réduire la fréquence d'échantillonnage de 48 kHz à 12.8 kHz. Cette dernière a été choisie afin de maximiser la bande utile tout en facilitant l'alignement des descripteurs audio et vidéo extraits. Pour réaliser ce sous-échantillonnage nous avons utilisé le logiciel *SMARC*[1] de Prado [2009] et développé à Télécom ParisTech.

1. SMart Audio Rate Converter - http ://audio-smarc.sourceforge.net/

Corpus *On n'a pas tout dit*

Le corpus *On n'a pas tout dit* a lui été constitué par Orange Labs, France Telecom dans le cadre de la thèse de Meriem Bendris (voir Bendris [2011]). Il s'agit de cinq émissions au format MPEG-1 et échantillonnées à 48 kHz (voir tableau A.2). Contrairement au corpus *Le Grand Échiquier*, celui-ci n'a pas nécessité de ré-échantillonnage de la piste audio.

VIDEO	
Dimensions	750 × 576 pixels
Codec	MPEG-1 video
Fréquence d'aquisition	25 images/seconde
Débit	~ 5500 kb/s
AUDIO	
Codec	MPEG-1 Audio, couche 2
Canaux	stéréo
Fréquence d'échantillonnage	48000 Hz
Débit	192 kb/s

Tableau A.2 – Propriétés des émissions du corpus *On n'a pas tout dit*.

Annexe B

Tableau récapitulatif des émissions du corpus *Le Grand Échiquier*

Nº	Identifiant notice	Date de diffusion	Titre	Durée
1	CPB76068458	27/05/1976	RAYMOND DEVOS	3h17mn
2	CPB79052393	31/05/1979	LINO VENTURA	3h03mn
3	CPB81052332	26/02/1981	JACQUES DUTRONC	3h08mn
4	CPB86006610	09/04/1986	ORCHESTRE NATIONAL DE FRANCE	3h09mn
5	CPB81050169	26/03/1981	MARIE PAULE BELLE	2h40mn
6	CPB81052012	24/09/1981	RUGGERO RAIMONDI	2h50mn
7	CPB81052234	25/06/1981	SERGE BAUDO ET L'ORCHESTRE DE LYON	3h00mn
8	CPB81054341	30/04/1981	LES MONDES DU COMMANDANT COUSTEAU	3h15mn
9	CPB82050299	27/01/1982	FREDERIC DARD ET SAN ANTONIO	2h40mn
10	CPB82050772	30/06/1982	ETIENNE VATELOT	2h57mn
11	CPB82051110	20/09/1982	ORCHESTRE ET CHŒURS DU THEATRE NATIONAL DE L'OPERA DE PARIS	3h00mn
12	CPB82051645	15/11/1982	UNE FLUTE EN OR POUR JEAN-PIERRE RAMPAL	2h40mn
13	CPB82053671	19/05/1982	LES FEMMES ET LES ENFANTS D'ABORD	2h45mn
14	CPB82053732	28/07/1982	JACQUES LAFFITE	2h40mn
15	CPB82054108	31/03/1982	ANGELO BRANDUARDI	2h37mn
16	CPB82055196	18/10/1982	MICHEL SARDOU	2h51mn
17	CPB83052569	20/06/1983	EN DIRECT DU TEP	3h00mn
18	CPB83052903	10/10/1983	COCTEAU VIVANT	4h30mn
19	CPB83053123	19/09/1983	CENTENAIRE DE L'ALLIANCE FRANCAISE	2h45mn
20	CPB83053882	21/11/1983	PLACIDO DOMINGO	3h30mn
21	CPB83055721	21/03/1983	C'EST LE PRINTEMPS DIDIER PIRONI	3h13mn
22	CPB83057057	18/04/1983	FRANCOIS RENE DUCHABLE PATRICK SEGAL	2h54mn
23	CPB84051413	25/06/1984	DANIEL BAREMBOIM	3h11mn
24	CPB84052346	15/10/1984	GERARD OURY	3h26mn
25	CPB84055197	19/11/1984	JESSYE NORMAN	3h15mn
26	CPB84056373	17/12/1984	PREMIERS DE CORDEE	3h34mn
27	CPB84057320	28/05/1984	L'INSTITUT CURTIS	2h45mn
28	CPB85050444	21/01/1985	SERGE LAMA - LES BATTANTS	3h04mn
29	CPB85100151	08/09/1985	BERNARD HINAULT	2h23mn
30	CPB85101320	22/09/1985	RAYMOND DEVOS OU L'ELOGE DE LA FOLIE	2h13mn
31	CPB85103721	15/04/1985	MICHEL JONASZ	3h13mn
32	CPB85104049	03/11/1985	MICHEL BERGER	2h17mn

Nº	Identifiant notice	Date de diffusion	Titre	Durée
33	CPB86004452	12/02/1986	VLADIMIR ASHKENASY	2h59mn
34	CPB86008274	11/06/1986	BARBARA HENDRICKS	2h56mn
35	CPB86011349	24/09/1986	TOUTE LA MUSIQUE POUR L'ARBRE DE VIE	3h29mn
36	CPB87001991	21/01/1987	HENRI SALVADOR	3h11mn
37	CPB87002568	11/03/1987	REGINE	2h51mn
38	CPB87004037	15/04/1987	LES SOLEILS DU PRINTEMPS	3h25mn
39	CPB87008572	09/09/1987	JEAN D'ORMESSON	2h55mn
40	CPB87011822	18/11/1987	MIREILLE MATHIEU	2h35mn
41	CPB87013464	23/12/1987	UNE NUIT A L'OPERA	3h09mn
42	CPB88000166	29/01/1981	TINO ROSSI	3h12mn
43	CPB88000401	13/01/1988	VIVE LA RENTREE	2h25mn
44	CPB88006265	11/05/1988	ISRAEL : QUARANTE ANS APRES	3h05mn
45	CPB88012790	24/10/1988	DEUX HOMMES DANS LE SIECLE	1h54mn
46	CPB89000622	26/12/1988	CAROLINE DE MONACO	3h08mn
47	CPB89002637	27/02/1989	DE MOSCOU A LENINGRAD	2h44mn
48	CPB89003871	27/03/1989	JULIA MIGENES	2h21mn
49	CPB89005217	24/04/1989	LE PRINTEMPS DE BEJART	2h22mn
50	CPB89005826	15/05/1989	FRANCIS HUSTER : 10 ANS AVANT L'AN 2000	2h28mn
51	CPB89009941	25/09/1989	TRENTE ANS APRES	2h15mn
52	CPB91005334	17/06/1985	UN SOIR AU QUEBEC	2h45mn
53	CPB8405214601	16/04/1984	SPORT	3h05mn
54	CPB8405343501	23/01/1984	CLAUDE BRASSEUR ET SES INVITES	3h15mn

Annexe C

Présentation explicative d'une notice documentaire

La notice présentée ici est typique de celles fournies avec le corpus *Le Grand Échiquier*. Ainsi les champs suivants ont été renseignés par le documentaliste chargé de fournir la notice documentaire :

- **Identifiant de la notice :** identifiant unique dans la base documentaire
- **Titre propre :** titre de l'émission. Par exemple « Raymond Devos »
- **Titre collection :** par exemple « *Le Grand Échiquier* »
- **Société de programmes :** nom de la chaîne, par exemple « A2 »
- **Diffusion :** date de diffusion de l'émission
- **Durée :** durée de l'émission
- **Descripteurs :** champ contrôlé par le thésaurus (descripteurs thématiques, descripteurs de lieux, descripteurs image, descripteurs son)
- **Nature de production :** par exemple : production propre, coproduction, achat de droits, etc.
- **Producteurs :** noms et rôles des producteurs
- **Genre :** par exemple : interview, documentaire, série, sketch, journal télévisé, etc.(51 genres possibles)
- **Thématique :** par exemple : spectacle, cinéma, ethnologie, faune, loisirs, variétés, etc. (46 thématiques possibles)
- **Générique :** réalisateur(s), présentateurs(s), participant(s) apparaissant au générique
- **Résumé :** résumé en texte libre
- **Œuvres :** œuvres référencées apparaissant dans l'émission
- **Corpus :** nom des corpus thématiques auxquels l'émission appartient
- **Statut de numérisation :** par exemple : numérisé, non numérisé, segmenté (avec time-codes), numérisé et en ligne, etc.
- **Document dévolu INA :** si les droits d'exploitation sont dévolus à l'Ina
- **Domaine :** par exemple : pour le site web de l'Ina

- **Date de création :** date de création de la notice
- **Date de modification :** date de modification de la notice
- **Documentaliste :** nom du documentaliste ayant créé la notice
- **Type de notice :** par exemple : notice isolée, notice sommaire (d'un journal télévisé par ex), notice sujet (id.)
- **Type de fonds :** parmi les différents fonds de l'Ina
- **Catalogage :** indique si le catalogage (titres et générique) est terminé
- **Indexation :** indique le niveau d'indexation et de validation atteint
- **Matériels :** lien vers les bases de référencement des matériels (supports physiques et informatiques)

Identifiant de la notice :	CPB76068458
Titre propre :	Raymond Devos
Titre collection (Aff.) :	Le grand échiquier
Société de programmes :	A2
Diffusion (aff.) :	27/05/1976 - type date: Diffusé -heure:20:29:00 - canal:2eme chaîne (A2)
Durée :	03:17:00
Nature de production :	Production propre
Producteurs (Aff.) :	Producteur ou co-producteur - Antenne 2 (A2) - Paris - 1976
Genre :	Spectacle TV ;
Thématique :	Variétés ;
Générique (Aff. Lig.) :	REA Flédérick, André ; PRE Chancel, Jacques ; INT Devos, Raymond ; INT Bernard, André ; INT Simon, Michel ; INT Gabillot, Eric (Danseur claquettes) ; INT Trankin, Daniel (Danseur claquettes) ; INT Trenet, Charles ; INT Cohen, Léonard ; INT Adamo, Salvatore ; INT Duvaleix, Christian (Fantaisiste) ; INT Bataillon de Joinville ; INT Marceau, Marcel ; INT Lewis, Jerry ; INT Fays, Raphaël (Guitariste) ; INT Davoust, Eric (Pianiste) ; INT Verstraete, Charles (Accordéoniste) ; INT Perrot, Jacques (Fantaisiste) ; INT Laure, Odette ; INT Grenier, Gilbert ; INT Vernes, Lucien (Prof. Piano) ; INT Laloux, Daniel (Tambour) ; INT Tambours de Saint Remy les Chevreuse ; PAR Rigerie, Stanislas (Prof. Piano) ;
Identifiant Matériels (info.) :	MGCPB0001696 . 01 (1) MGCPB0001696 . 02 (2) MGCPB0001696 . 03 (3) MGCPB0029110 . 01 MGCPB0029110 . 02 MGCPB0029110 . 03 MGCPB0031254 . 01 MGCPB0031254 . 02 MGCPB0031254 . 03 MGCPB0278910 . 01 MGCPB0278910 . 02 MGCPB0278910 . 03
Nom fichier segmenté (info) :	MGCPB0001696.01_000300_013435.mps /MGCPB0001696.02_000002_013346.mps /MGCPB0001696.03_000000_002320.mps /
Résumé :	Invite de l'émission, Raymond DEVOS joue du piano puis du tuba, explique comment il a commence a jouer du tuba, parle du rôle de l'artiste, de ses monologues, de la préparation de ses sketchs, de Michel SIMON, de sa superstition, du mime et interprète plusieurs sketches :la 1ere fois ou il s'est présente au public, répétition orchestre dans ring (doc. Archives), le plaisir des sens, le fils d'ABRAHAM, l'horoscope, l'agent de police, l'hôtelier, la valise, avec le mime Marceau "les mains", le cri d'alarme et avec Daniel LALOUX "les tambours". Il essaie de danser des claquettes. Chante "bonsoir jolie madame" et "les bohémiens" de Félix LECLERC. Raconte histoire du piano DEVOS puis joue thème musical tout en l'expliquant. André Bernard joue concerto pour trompette de Haydn. Extrait grand échiquier BEART : interview Michel SIMON et chanson. Éric GABILLOT et Daniel FRANKIN dansent tour a tour des claquettes. Ch. Trenet chante "cinq ans de marine", "vive la vie, vive l'amour" et "le soleil et la lune". Léonard COHEN "Suzanne", "lover, lover" et en français et en anglais "bird on my wire". Int. sur ses chansons. Adamo chante "grandis pas mon fils". Numéro de la coquille Saint Jacques par Christian DUVALEX. M. MARCEAU mime "la révolte de l'automate". Jerry LEWIS il y a 15 ans dans mime du PDG et même mime en parallèle tourne aujourd'hui. Raphaël FAYS joue plusieurs morceaux a la guitare.

Eric DAVOUST interprète "vol du bourdon". Stanislas RIGERIE raconte les cours de piano qu'il a donnes à R. DEVOS. Charles VERSTRAETE joue accordéon. Jack PERROT interprète musique moderne avec la bouche et les doigts.
Odette LAURE "ça tourne pas rond dans ma p'tite tète". M. MARCEAU mime "le bien et le mal". G. GRENIER joue guitare et chante, Lucien VERNES parle de son élève R. DEVOS. Final avec batterie des tambours de SAINT REMY les Chevreuse.

Oeuvres :

RAYMOND DEVOS interprète : "Le plaisir des sens", "Le fils d'Abraham", "L'horoscope", "L'agent de police", "L'hôtelier", "L'homme qui fait sa valise", "Les mains", "Le cri d'alarme", "Les tambours", "Le clou", "Le bout du bout".

- RAYMOND DEVOS chante : "Bonsoir jolie Madame", "Les bohémiens".
- ANDRE BERNARD joue le concerto pour trompette de Haydn. CHARLES TRENET chante : "Cinq ans de marine", "Vive la vie, vive l'amour", "Le soleil et la lune".
- LEONARD COHEN chante : "Suzanne", "Lover, lover", "Bird on my wire".
- ADAMO chante "Grandis pas mon fils".
- MARCEAU mime : "La révolte de l'automate", "Le bien et le mal".
- ERIC DAVOUST interprète : "Le vol du bourdon".
- ODETTE LAURE : "C'a tourne pas rond dans ma petite tête".

Corpus (Aff.) : Corpus: Réalisations émissions intégrales par André FLEDERICK - (Corpus INA > PERSONNALITE > DEVOS RAYMOND > Portraits et interviews > Réalisations émissions intégrales par André FLEDERICK)

Statut de numérisation : Numérisé/segmenté avec matériel de pré-sélection
Document dévolu INA : Dévolu INA
Document fonds TF1 : Non
Domaine (Aff.) : OGP (Offre Grand Public)- X (X) - Pas en ligne

Date de création : 02/06/1976
Date de modification : 19/09/2006
Documentaliste : MAC
Dernier intervenant : FER
Thèque : CP (Vidéothèque production)
Type de notice : Notice isolée
Type de fonds : Production
Classe de niveau : Défaut Production
Anciens Supports : MGTO 2',COULEUR : PROD. 41235
Catalogage (info.) : Atteint: Oui Validé: O Date: 19/09/2006
Indexation (info.) : Atteint: Oui Validé: N
Matériels (Détail) : [MSV] - MGCPB0001696 . 01/03 - Stat. Num.: Numérisé et en ligne - Rang: 1 - TC IN: 00:03:00:09 TC OUT: 01:34:35:19 - Filiation: MGCPB0031254
Format: 1/2 BSP - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: MASTERUN - Durée: 00:00:00 -
Stat.Vers.: Versé (13/03/2008) - Localisation: CPESSU4 (13/03/2008) - N° stock.: MSV136492 - N° bande: 0001696 - N°EM: PE41235

[MSV] - MGCPB0001696 . 02/03 - Stat. Num.: Numérisé et en ligne - Rang: 2 - TC IN: 00:00:02:11 TC OUT: 01:33:46:03 - Filiation: MGCPB0031254
Format: 1/2 BSP - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: MASTERUN - Durée: 00:00:00 -
Stat.Vers.: Versé (13/03/2008) - Localisation: CPESSU4 (13/03/2008) - N° stock.: MSV136493 - N° bande: 0001697 - N°EM: PE41235

[MSV] - MGCPB0001696 . 03/03 - Stat. Num.: Numérisé et en ligne - Rang: 3 - TC IN: 00:00:00:15 TC OUT: 00:23:20:08 - Filiation: MGCPB0031254
Format: 1/2 BSP - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: MASTERUN - Durée: 00:00:00 -
Stat.Vers.: Versé (13/03/2008) - Localisation: CPESSU4 (13/03/2008) - N° stock.: MSV136494 - N° bande: 0001698 - N°EM: PE41235

[SRC] - MGCPB0029110 . 01/03 -
Format: 2P HB - Définition: 625 - Format img.: 4/3 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: PA - Durée: 01:28:00 - Etat supp.: 3
Stat.Vers.: Versé (31/12/1976) - Localisation: CPESSU1 (22/05/2009) - N° stock.: 243 - N° bande: 0029110 - N°EM: PE41235

[SRC] - MGCPB0029110 . 02/03 -
Format: 2P HB - Définition: 625 - Format img.: 4/3 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: PA - Durée: 01:28:00 - Etat supp.: 3
Stat.Vers.: Versé (31/12/1976) - Localisation: CPESSU1 (22/05/2009) - N° stock.: 230 - N° bande: 0029012 - N°EM: PE41235

[SRC] - MGCPB0029110 . 03/03 -
Format: 2P HB - Définition: 625 - Format img.: 4/3 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: PA - Durée: 01:28:00 - Etat supp.: 3
Stat.Vers.: Versé (31/12/1976) - Localisation: CPESSU1 (22/05/2009) - N° stock.: 613 - N° bande: 0031531 - N°EM: PE41235

[COM] - MGCPB0278910 . 02/03 - Filiation: MGCPB0031254
Format: 1P C - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: COP - Durée: 00:00:00 - Etat supp.: 3
Stat.Vers.: Détruit Localisation: Détruit (06/04/2004) - N° bande: 0278911 - N°EM: PS41235

[COM] - MGCPB0278910 . 03/03 - Filiation: MGCPB0031254
Format: 1P C - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: COP - Durée: 00:00:00 - Etat supp.: 3
Stat.Vers.: Détruit Localisation: Détruit (06/04/2004) - N° bande: 0278912 - N°EM: PS41235

Matériels dispo (Détail) :

[MSV] - MGCPB0001696 . 01/03 - Stat. Num.: Numérisé et en ligne - Rang: 1 - TC IN: 00:03:00:09 TC OUT: 01:34:35:19 - Filiation: MGCPB0031254
Format: 1/2 BSP - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: MASTERUN - Durée: 00:00:00 -
Stat.Vers.: Versé (13/03/2008) - Localisation: CPESSU4 (13/03/2008) - N° stock.: MSV136492 - N° bande: 0001696 - N°EM: PE41235

[MSV] - MGCPB0001696 . 02/03 - Stat. Num.: Numérisé et en ligne - Rang: 2 - TC IN: 00:00:02:11 TC OUT: 01:33:46:03 - Filiation: MGCPB0031254
Format: 1/2 BSP - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: MASTERUN - Durée: 00:00:00 -
Stat.Vers.: Versé (13/03/2008) - Localisation: CPESSU4 (13/03/2008) - N° stock.: MSV136493 - N° bande: 0001697 - N°EM: PE41235

[MSV] - MGCPB0001696 . 03/03 - Stat. Num.: Numérisé et en ligne - Rang: 3 - TC IN: 00:00:00:15 TC OUT: 00:23:20:08 - Filiation: MGCPB0031254
Format: 1/2 BSP - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: MASTERUN - Durée: 00:00:00 -
Stat.Vers.: Versé (13/03/2008) - Localisation: CPESSU4 (13/03/2008) - N° stock.: MSV136494 - N° bande: 0001698 - N°EM: PE41235

[FAB] - MGCPB0031254 . 03/03 -
Format: 2P HB - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: BDE - Durée: 00:00:00 - Etat supp.: 4
Stat.Vers.: Versé (01/09/1992) - Localisation: CPESSU1 (04/06/2009) - N° stock.: 654 - N° bande: 0031993 - N°EM: PE41235

[FAB] - MGCPB0031254 . 02/03 -
Format: 2P HB - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: BDE - Durée: 00:00:00 - Etat supp.: 4
Stat.Vers.: Versé (01/09/1992) - Localisation: CPESSU1 (04/06/2009) - N° stock.: 603 - N° bande: 0031478 - N°EM: PE41235

[FAB] - MGCPB0031254 . 01/03 -
Format: 2P HB - Définition: 625 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: BDE - Durée: 00:00:00 - Etat supp.: 4
Stat.Vers.: Versé (01/09/1992) - Localisation: CPESSU1 (04/06/2009) - N° stock.: 560 - N° bande: 0031254 - N°EM: PE41235

[SRC] - MGCPB0029110 . 03/03 -
Format: 2P HB - Définition: 625 - Format img.: 4/3 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: PA - Durée: 01:28:00 - Etat supp.: 4
Stat.Vers.: Versé (31/12/1976) - Localisation: CPESSU1 (22/05/2009) - N° stock.: 613 - N° bande: 0031531 - N°EM: PE41235

[SRC] - MGCPB0029110 . 02/03 -
Format: 2P HB - Définition: 625 - Format img.: 4/3 - Son: MONO - Coul.: COULEUR -
Filière: MG - Type Mat.: PA - Durée: 01:28:00 - Etat supp.: 4
Stat.Vers.: Versé (31/12/1976) - Localisation: CPESSU1 (22/05/2009) - N° stock.: 230 - N° bande: 0029012 - N°EM: PE41235

Annexe D

Détails sur le thésaurus de l'Ina

Le thésaurus de l'Ina est organisé en neuf axes thématiques correspondant à neuf parties, où chacun des termes qui les compose est organisé d'une manière hiérarchique [1]. Par exemple : le mot clé porte-avions est « fils de » bateau de guerre, lui même « fils de » matériel de marine, lui même « fils de » marine, elle même « fils de » armée, terme rattaché à la rubrique Politique Française. Afin de pouvoir classer chacun des sujets sous une rubrique, les mots du thesaurus ont été assimilés à un plan de classement et chaque descripteur a été associé à une rubrique de manière univoque. Ainsi une table de correspondance a été créée entre les mots clés du thesaurus et les rubriques permettant le classement d'un sujet sous la rubrique la plus pertinente. Un sujet est classé sous une rubrique et une seule. Il faut souligner ici que le langage d'un thésaurus est codé et formalisé.

Par exemple : pour le porte avions Clemenceau, tous les sujets ayant trait à son activité en tant que bateau de guerre de la Marine française sont classés sous la rubrique Politique française. Pour indexer les sujets traitant de son désamiantage les mots « décontamination » et « environnement » sont utilisés. Ils sont associés à la rubrique Environnement. Cependant sont classés systématiquement sous la rubrique International les sujets qui se déroulent dans un pays autre que la France, relatifs à leur politique intérieure (élections, crise politique...) ou liés au terrorisme, aux conflits, à la géopolitique... Enfin, au thésaurus général s'ajoutent deux parties complémentaires : les termes géographiques et la liste des personnalités. Les quatorze rubriques du thésaurus de l'Ina sont les suivantes :

- **Catastrophe :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié par exemple : aux catastrophes, aux catastrophes naturelles, aux accidents, mais aussi à celui des opérations de sauvetage...
- **Culture-loisirs :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié par exemple : aux arts, aux arts plastiques, au cinéma, au cirque, à la littérature, au théâtre et aux loisirs comme la photographie, les fêtes ou encore les festivals...
- **Économie :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux grands secteurs économiques comme l'agriculture et l'élevage, l'industrie, le tourisme, les transports et aussi celui du monde de la finance...
- **Éducation :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, à l'enseignement et à la formation permanente...

1. source : http ://www.ina-sup.com/ressources/methodologie

- **Environnement :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, au climat, à l'écologie, aux éléments naturels, à l'urbanisme...
- **Faits divers :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux délits comme les vols et les agressions...
- **Histoire Hommage :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux anniversaires, aux commémorations, à l'histoire et les rétrospectives...
- **International :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux conflits armés ou pas, au terrorisme, à la géopolitique, à l'Union européenne et la vie de politique intérieure de tous les pays autres que la France.
- **Justice :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux procédures judiciaires, à la réforme de la justice, aux prisons...
- **Politique française :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, à la politique intérieure, les différentes élections, la sécurité du territoire, la police, la défense nationale, l'armée...
- **Santé :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux maladies et thérapeutiques, à l'éthique...
- **Sciences et techniques :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux sciences dures, à la conquête de l'espace, la vulgarisation...
- **Société :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, à la politique sociale, à l'action sociale, à la démographie, à la religion, aux conditions de vie, au travail...
- **Sport :** sous cette rubrique sont classés les documents indexés avec le vocabulaire du thesaurus lié, par exemple, aux disciplines sportives, au dopage, aux manifestations sportives...

Annexe E

Présentation du protocole d'annotation

La campagne d'annotation menée pendant l'été 2009 a nécessité la mise en place d'un protocole très détaillé. Les champs à renseigner par l'annotateur on été définis comme cela est spécifié dans le tableau E.1. Pour tous les événements l'annotateur a donné les temps de début et fin (*timecodes*) et parfois des informations complémentaires. Ainsi, pour les performances musicales et non-musicales devaient être indiqués les identités des interprètes, œuvres et genre (musique classique, numéro de cirque, lecture, etc.). Pour les rires et applaudissements, l'annotateur devait spécifier si ils étaient le fait du public ou de participants présents sur le plateau. Des événements du type lancement (généralement réalisé par le présentateur Jacques Chancel), arrivée d'un nouvel invité, émotion (fou-rires, pleurs, recueillement, etc.) ou mouvement brusque ont également été répertoriés. Des considérations de montage — par exemple les plans de coupe (ici plongées et vue du public) — de même que la localisation de l'action montrée à l'écran ont également été prises en compte.

Événement à détecter	**Timecodes**	**Genre**	**Identité**	**Œuvre**	**Public - participants**	**N°**
performance musicale	✓	✓	✓	✓		
performance non-musicale	✓	✓	✓	✓		
insert	✓	✓				
parole	✓					
applaudissements	✓				✓	
rires	✓				✓	
nom d'intervenant	✓		✓			
arrivée nouvel invité	✓		✓			
lancement	✓		✓			
émotion	✓		✓			
mouvement brusque	✓		✓			
1 gros plan par participant	✓		✓			
1 plan moyen par participant	✓		✓			
localisation - scène	✓					
localisation - plateau	✓					
plan de coupe - plongée	✓					
plan de coupe - vue du public	✓					
alignement résumé notice	✓					✓

Tableau E.1 – Présentation des différentes catégories d'événements audiovisuels annotées pour le corpus *Le Grand Échiquier*

L'annotateur a également dû trouver un gros plan et un plan moyen de chaque intervenant de l'émission, ces intervenants étant listés dans le champ *générique* de la notice documentaire associée. Enfin, le champ *résumé* de chaque notice a subi un pré-traitement automatique pour identifier et numéroter de façon claire tous les événements d'importance notés par le documentaliste. Chaque phrase a ainsi été divisée en propositions identifiées par un numéro pour que l'annotateur puisse isoler chacune d'entre elles et retrouver les *timecodes* de début et de fin correspondants. Un exemple de ce pré-traitement pour le champ résumé de la notice présentée dans l'annexe C est donné plus bas :

1 : Invité de l'émission, Raymond Devos joue du piano
2 : puis du tuba
3 : explique comment il a commencé à jouer du tuba
4 : parle du rôle de l'artiste
5 : de ses monologues
6 : de la préparation de ses sketches
7 : de Michel Simon
8 : de sa superstition
9 : du mime
10 : et interprète plusieurs sketches :
11 : « la 1ere fois ou il s'est présenté au public »
12 : « répétition orchestre dans ring » (doc archives)
13 : « le plaisir des sens »
14 : « le fils d'Abraham »
15 : « l'horoscope »
16 : « l'agent de police »
17 : « l'hôtelier »
18 : « l'homme qui fait sa valise »
19 : avec le mime Marceau « les mains »
20 : « le cri d'alarme »
21 : et avec Daniel Laloux « les tambours »
22 : Il essaie de danser des claquettes
23 : chante « bonsoir jolie madame »
24 : et « les bohemiens » de Félix Leclerc
25 : raconte l'histoire du piano Devos
26 : puis joue un thème musical tout en l'expliquant
27 : André Bernard joue le concerto pour trompette de Haydn
28 : Extrait Grand Échiquier Béart : interview Michel Simon et chanson
29 : Éric Gabillot et Daniel Frankin dansent tour à tour des claquettes
30 : Charles Trenet chante « cinq ans de marine »
31 : « vive la vie, vive l'amour »
33 : « le soleil et la lune »34 : Leonard Cohen « suzanne »
35 : « lover lover »
36 : et en francais et en anglais « bird on my wire »
37 : interview sur ses chansons
38 : Adamo chante « grandis pas mon fils »
39 : Numéro de la coquille Saint-Jacques par Christian Duvalex
40 : Le mime Marceau mime « la révolte de l'automate »
41 : Jerry Lewis il y a 15 ans dans mime du PDG
42 : et même mime en parallèle tourné aujourd'hui

Annexe F

Instructions et questionnaire destinés aux utilisateurs

Dans l'expérience utilisateur présentée au chapitre 2, vingt sujets ont testé les avantages de la méthode de structuration que nous proposons dans cette thèse. Nous reproduisons dans cette annexe les documents qu'ils ont eu à leur disposition pour réaliser cette expérience. Le premier de ces documents est une présentation du protocole expérimentale ainsi que du logiciel ELAN utilisé. Le second document était à remplir à l'issue du test. Il s'agit d'un questionnaire interrogeant les utilisateurs sur leurs impressions suite à l'expérience et sur l'intérêt à leurs yeux du schéma de structuration proposé.

INSTRUCTIONS POUR LA RECHERCHE D'EXTRAITS AUDIOVISUELS DU GRAND ECHIQUIER

Cette tâche consiste à retrouver temporellement des extraits audiovisuels correspondants à des éléments de notices identifiés par des documentalistes de l'Ina (Institut National de l'Audiovisuel) pour quatre émissions du talk show *« Le Grand échiquier »*. Il y a quatre éléments à identifier par émission.
Pour cela deux scénarii sont proposés (deux émissions pour chacun) :

- alignement réalisé par la navigation avec les fonctionnalités de base du *player* vidéo (lecture, pause, avance/retour rapide, *slider*, etc.)
- alignement réalisé par la navigation dans des champs d'annotation prédéfinis obtenus de façon automatique en plus des fonctionnalités de base du *player* vidéo

L'utilisateur se servira du programme ELAN pour la visualisation/navigation dans l'émission à annoter et d'une interface graphique MATLAB pour entrer les résultats de recherche.

L'expérience est divisée en deux : une séquence d'entrainement afin de se familiariser à la manipulation du logiciel ELAN puis une séquence de test.

1. Interface ELAN

a. A TELECOM ParisTech (Linux)

- ouvrir un terminal et taper :
- $ source /opt/tsi/devel/elan-3.9.0_linux32_i686.bashrc
- $ elan
- dans le programme ELAN sélectionner Options/Langue/Français
- Fichier/Ouvrir et dans sélectionner /Desktop/EXPERIENCE/1_train/ ou /Desktop/EXPERIENCE/1_test/ sélectionner le fichier train_X_A.eaf / train_X_B.eaf ou test_X_A.eaf / test_X_B.eaf suivant le scénario choisi
- ELAN demande ensuite de localiser le fichier train_X.MPG ou test_X.MPG correspondant (celui-ci se trouve dans le même dossier)

b. A l'Ina (Windows)

- Aller, suivant l'expérience dans le dossier \Bureau\EXPERIENCE\1_train\ ou \Bureau\EXPERIENCE\1_test\
- Ouvrir le fichier train_X_A.eaf / train_X_B.eaf ou test_X_A.eaf / test_X_B.eaf suivant le scénario choisi
- ELAN demande ensuite de localiser le fichier train_X.MPG ou test_X.MPG correspondant (celui-ci se trouve dans le même dossier)

Suivant le scenario choisi, le logiciel ELAN doit afficher quelque chose de similaire à la figure 1 (pas de champs prédéfinis) ou la figure 2 (présence de champs prédéfinis)

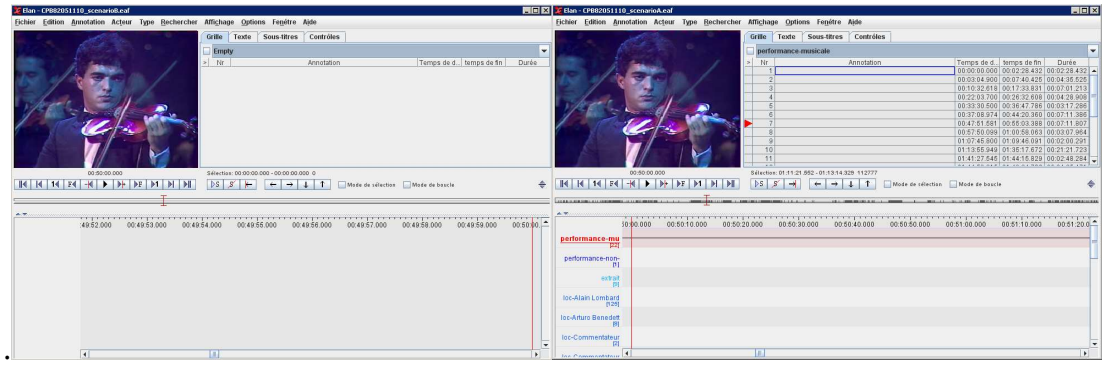


Fig. 1 et 2 *Apparence de l'interface ELAN interface pour l'annotation suivant le scenario*

c. *Éléments de navigation de base avec ELAN*

Fig. 3 *Fonctionnalités de navigation*

De gauche à droite de la figure 3, les fonctionnalités des 11 boutons sont:
- se déplacer au début de la vidéo
- se déplacer au curseur précédent (quelques minutes en général)
- se déplacer une seconde avant
- se déplacer une trame avant
- se déplacer un dixième de seconde avant
- play/pause
- se déplacer un dixième de seconde après
- se déplacer une trame après
- se déplacer une seconde après
- se déplacer au curseur suivant (quelques minutes en général)
- se déplacer à la fin de la vidéo

L'utilisateur peut utiliser le *slider* temporel pour naviguer plus grossièrement dans la vidéo.

***Fig. 4** Slider temporel*

Enfin, l'utilisateur peut changer la précision de l'affichage en effectuant un clic droit sur la ligne de temps comme sur la figure 5.

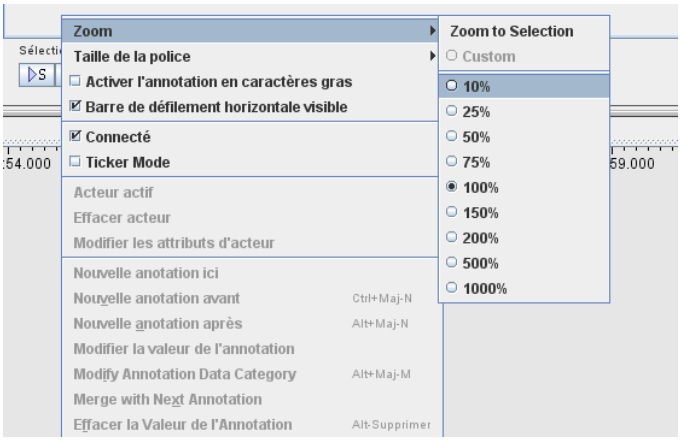


***Fig. 5** Zoom*

d. Éléments de navigation dans les champs prédéfinis avec ELAN

Si l'utilisateur teste la recherche d'extraits par champs prédéfinis, alors il peut naviguer également dans les champs pré-annotés (cf figure 6).

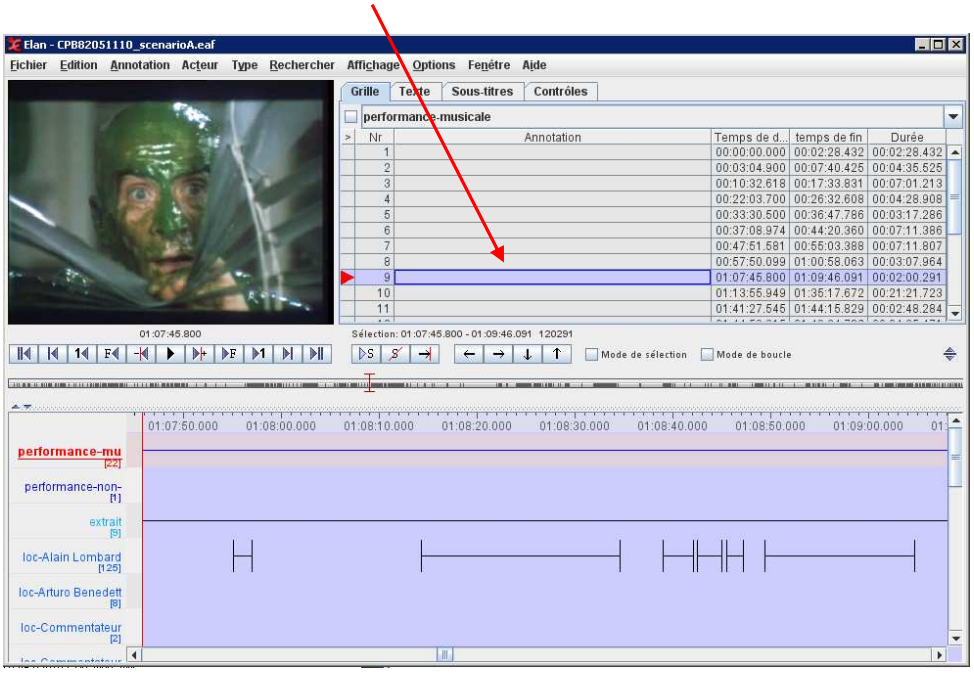


***Fig. 6** Navigation à l'aide des champs prédéfinis*

Dans l'onglet grille il suffit de sélectionner le champ qui intéresse l'utilisateur puis de cliquer sur une occurrence. Le curseur temporel se place alors automatiquement au début de l'extrait. L'utilisateur peut également utiliser la partie basse de l'interface pour la navigation comme indiqué dans la figure 7 :

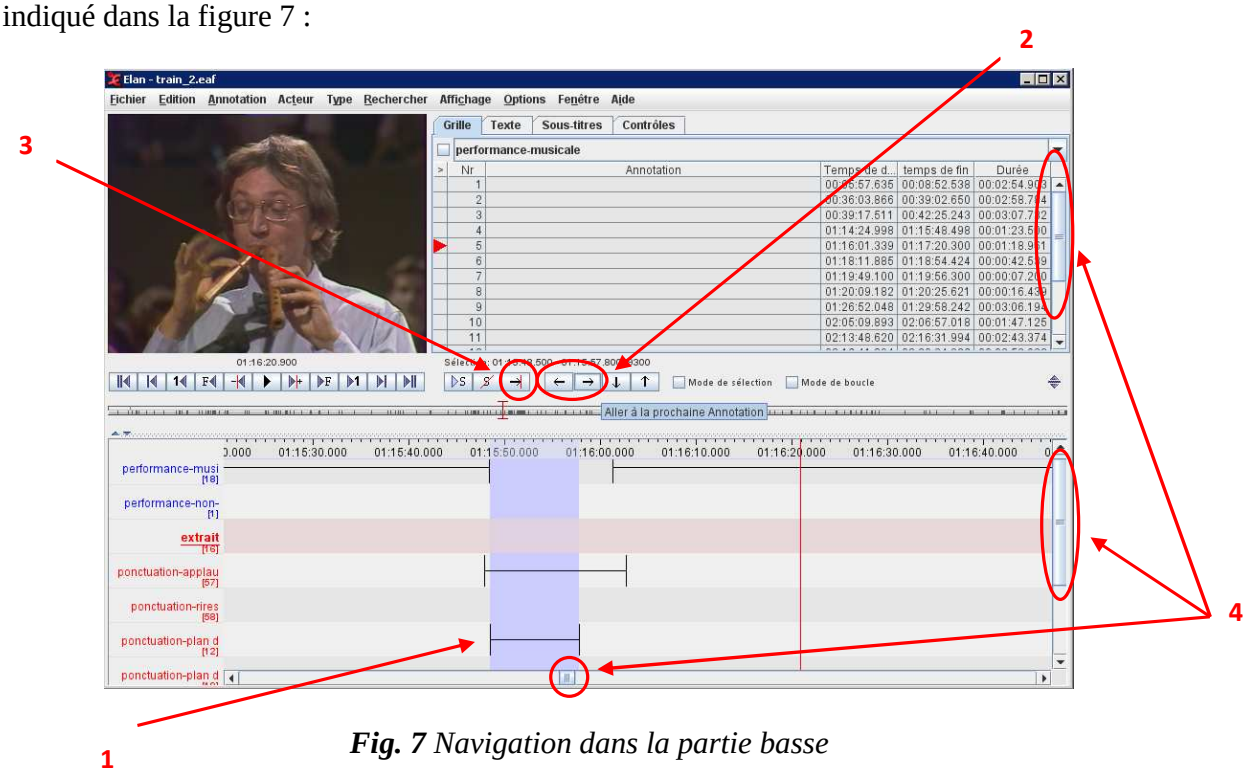


Fig. 7 *Navigation dans la partie basse*

Ainsi, un simple clic sur un segment de la partie basse sélectionne tout le segment en bleu (1). On peut ensuite se déplacer dans les annotations précédentes avec les boutons (2). Le bouton (3) replace le curseur au début ou à la fin du segment sélectionné. Enfin, de nombreux « ascenseurs » permettent de naviguer dans les annotations.

Les champs prédéfinis sont de natures diverses. Ils constituent des éléments de structure de base. Le tableau 1 offre un récapitulatif des ces champs avec le code de couleur utilisé :

éléments de structure		
contenu	performance	musicale
		non-musicale
	extrait	photo
		film/reportage
	parole	locuteurs
ponctuation	audio	applaudissements
		rires
	vidéo	plan de coupe
location	scène	-
	plateau-interview	-
	extérieur	-

Tableau 1 *Description des éléments de structure disponibles à la navigation*

Le champ contenu est divisé en 3 groupes constituant le corps du talk show : performance, extrait et parole, eux-mêmes subdivisés. A l'intérieur du champ parole, le champ locuteur regroupe les interventions orales de chaque personne au cours de l'émission (ex : Jacques Chancel, Coluche, etc.). Ces derniers sont classés dans l'ordre alphabétique (par prénoms).
Le champ ponctuation regroupe les éléments qui marquent des ruptures dans le déroulement de l'émission. Ainsi, les applaudissements indiquent généralement la fin d'une performance, les rires, la chute d'un gag, les plans de coupe, un changement de contenu, etc.
Enfin, le champ location indique où se déroule l'action montrée au téléspectateur. Celle-ci peut être sur le plateau d'interview, sur scène ou encore à l'extérieur du studio comme dans le cas d'extraits ou de duplex.
Il conviendra à l'utilisateur de se servir des champs lui paraissant les plus appropriés pour retrouver l'extrait demandé.

2. Interface MATLAB

Les éléments à retrouver dans la notice, en indiquant les segments temporels où ils se trouvent, sont affichés dans une interface graphique MATLAB. Pour la lancer, exécuter le les scripts train_fr.m et test_fr.m dans la console MATLAB. L'interface se présente comme indiqué dans la figure 6.

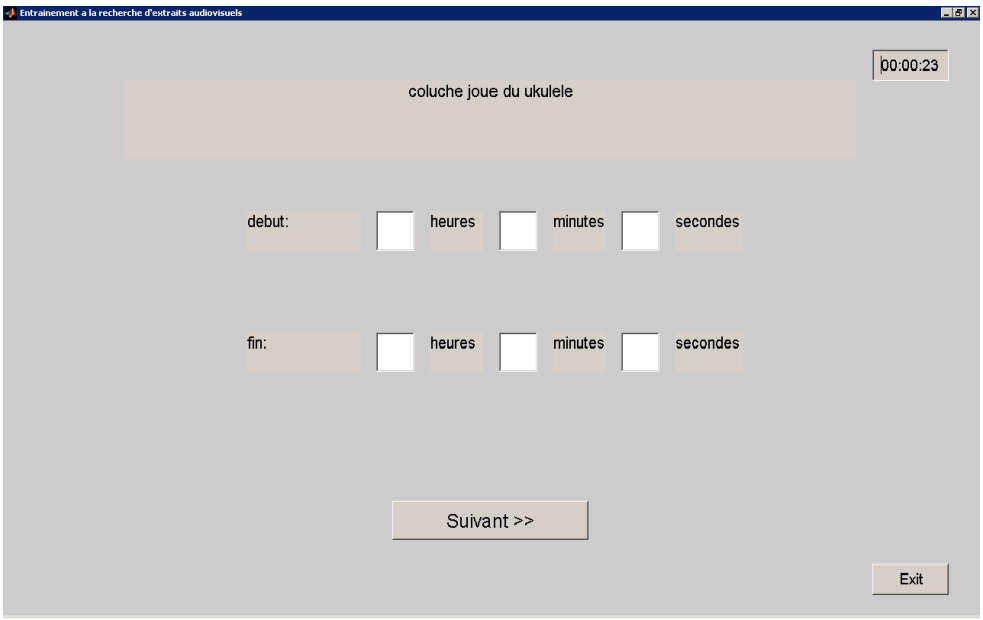


Fig. 6 *Interface MATLAB pour l'identification des segments temporels*

En haut, est indiqué à l'utilisateur l'extrait à retrouver dans l'émission. Il doit entrer des valeurs au format heure/minute/seconde pour le début et la fin avec une tolérance de 5 sec.

Ce qui est mesuré lors de cette expérience est la pertinence de l'outil proposé par la rapidité à retrouver les extraits. De fait, l'utilisateur doit apporter une méticulosité mesurée lors de la recherche des segments (d'où la tolérance de 5 sec).

3. Retour d’expérience

A l’issue de l’expérience, l’utilisateur devra remplir le questionnaire Retour_utilisateur.doc afin de connaître ses impressions sur l’outil proposé. Il devra le sauver sous la forme Retour_utilisateur_NOM.doc

4. Exemples d’extraits à retrouver

Voici le type d’extraits que vous aurez à retrouver dans les émissions :

- « jacques dutronc chante "croyez-vous que je sois jaloux?" »
- « laurent moriot fait un numero de claquettes »
- « jacques laffite evoque le championnat du monde de football »
- « les membres du quatuor ivaldi parlent de leur amitie tous les 4 »
- etc.

RETOUR UTILISATEUR SUR L'OUTIL DE RECHERCHE D'EXTRAITS

Ce questionnaire a pour but d'évaluer l'apport de la structuration par champs prédéfinis pour une tâche de recherche d'extraits audiovisuels.

Nom de l'utilisateur :

1 Avez-vous apprécié l'expérience ?
oui, tout à fait ☐ oui, plutôt ☐ non, plutôt pas ☐ non, pas du tout ☐

2 Si non, pour quelles raisons ?
trop long ☐ utilisation du logiciel ELAN délicate ☐ invités des émissions inconnus ☐
précisions si autre :

3 Aviez-vous déjà utilisé le logiciel ELAN auparavant ?
oui ☐ non, jamais ☐

4 Avez-vous trouvé l'utilisation d'ELAN difficile ?
oui, tout à fait ☐ oui, plutôt ☐ non, plutôt pas ☐ non, pas du tout ☐

5 Avez-vous eu la sensation de vous améliorer au maniement d'ELAN au cours de l'émission ?
oui, tout à fait ☐ oui, plutôt ☐ non, plutôt pas ☐ non, pas du tout ☐

6 Pensez-vous avoir été plus rapide avec ou sans l'utilisation des champs prédéfinis ?
plus rapide avec ☐ plus rapide sans ☐ ne sais pas ☐

7 Indépendamment du maniement d'ELAN, l'utilisation de champs prédéfinis vous parait-elle avantageuse pour la recherche d'extraits ?
oui, tout à fait ☐ oui, plutôt ☐ non, plutôt pas ☐ non, pas du tout ☐

8 Les champs prédéfinis suivants vous ont-ils paru utiles ?

- performance musicale	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- performance non musicale	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- extrait	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- locuteur	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- ponctuation applaudissements	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- ponctuation rires	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- ponctuation plan de coupe	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- location scène	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- location extérieur	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐
- location plateau	très utile ☐	assez utile ☐	peu utile ☐	inutile ☐

Annexe G

Algorithmes de calcul de flux optique et d'extraction de points d'intérêt (Kanade-Lucas-Tomasi)

Algorithme de calcul du flux optique

La plupart des méthodes d'estimation du flux optique reposent sur une hypothèse fondamentale : l'intensité lumineuse I se conserve entre deux images successives. Celle-ci peut être écrite sous la forme suivante pour deux dimensions x et y et en fonction du temps t :

$$I(x, y, t) = I(x + \delta x, y + \delta y, t + \delta t) \tag{G.1}$$

À l'aide d'un développement de Taylor au premier ordre on obtient :

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \tag{G.2}$$

Avec l'équation (G.1) on obtient alors :

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0 \tag{G.3}$$

soit,

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} = 0 \tag{G.4}$$

et donc,

$$\nabla I^T \cdot \vec{V} + I_t = 0 \tag{G.5}$$

avec $\nabla I = \left[I_x, I_y\right]^T$ le gradient spatial de l'intensité lumineuse, I_t la dérivée temporelle et $\vec{V}$ la vitesse recherchée.

Cette contrainte est néanmoins insuffisante pour determiner le flux complet $\vec{V}$. La méthode de Lucas et Kanade [1981] suppose que la vitesse locale est constante sur un voisinage spatial Ω, on minimise alors la fonctionnelle :

$$J_{LK} = \sum_{\Omega} W_1^2 \left[\nabla I^T \cdot \vec{V} + I_t\right]^2 \tag{G.6}$$

W_1 étant une fenêtre locale et pouvant également être interprétée comme la pondération du critère des moindres carrés. On donne généralement une importance plus grande au pixel central (filtrage de type gaussien).

Algorithme d'extraction de points d'intérêt

Les méthodes de détection de points d'intérêt (voir Harris et Stephens [1988] et Shi et Tomasi [1994] se basent sur la différence E d'intensité lumineuse dans un voisinage réduit dont l'un est décallée de $(\delta x, \delta y)$ par rapport à l'autre (pour deux dimensions) :

$$E(\delta x, \delta y) = \sum_{\delta x, \delta y} W_2(\delta x, \delta y) \left[(x+\delta x, y+\delta y) - I(x,y)\right]^2 \tag{G.7}$$

W_2 étant une fenêtre locale. En utilisant à nouveau un développement de Taylor au premier ordre on obtient :

$$I(x+\delta x, y+\delta y) = I(x,y) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y \tag{G.8}$$

Injecté dans l'équation G.7 on obtient alors :

$$E(\delta x, \delta y) = \sum_{\delta x, \delta y} W_2(\delta x, \delta y)\left(\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y\right)^2 \tag{G.9}$$

qui peut être réécrite comme :

$$E(\delta x, \delta y) = (\delta x, \delta y) A (\delta x, \delta y)^T \quad \text{avec} \quad A = \sum_{\delta x, \delta y} W_2(\delta x, \delta y) \begin{pmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{pmatrix} = \begin{pmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_x I_y \rangle & \langle I_y^2 \rangle \end{pmatrix} \tag{G.10}$$

A est la matrice de Harris et les symbôles $\langle\ \rangle$ dénotent un moyennage. Un point d'intérêt est caractérisé par une importante variation de E dans les directions du vecteur $(\delta x, \delta y)$.

Pour le détecteurs de points d'intérêt de Harris et Stephens [1988], tous les points pour lesquels le score R est supérieur à un seuil fixé sont considérés d'intérêt. Ce score est calculé suivant :

$$R = \det A - \kappa(\text{trace}\, A)^2 \tag{G.11}$$

avec κ à fixer empiriquement, $\det A = \lambda_1 \lambda_2$, $\text{trace}\, A = \lambda_1 + \lambda_2$, λ_1 et λ_2 étant les valeur propres de A.
Pour le détecteur de points d'intérêt de Shi et Tomasi [1994], le score R d'un point considéré est calculé suivant :

$$R = \min(\lambda_1, \lambda_2) \tag{G.12}$$

En effet, sous certaines hypothèses, les points d'intérêt détectés sont alors plus stables pour le suivi. C'est cette dernière méthode qui est implémentée par défaut dans le logiciel *OpenCV* (voir Bradski et Kaehler [2008]).

Annexe H

Classification par machine à vecteurs de support (SVM)

Dans les problèmes bi-classes, l'algorithme de classification par SVM recherche l'hyperplan qui sépare au mieux les exemples d'apprentissage appartenant aux différentes classes, c'est-à-dire avec un maximum de marge. L'hyperplan à maximum de marge est celui pour lequel la distance au point le plus proche de chaque côté est maximisé comme cela est montré sur la figure H.1. Avec un ensemble d'apprentissage $\mathscr{D}$ de n points et y_i l'étiquette du vecteur de caractéristiques x_i,

$$\mathscr{D} = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, +1\}\}_{i=1}^{n} \tag{H.1}$$

La solution à ce problème est l'hyperplan $\mathscr{H}$ d'équation (H.2) et qui vérifie l'équation (H.5)

$$w.x + b = 0 \quad \text{avec} \quad w \in \mathbb{R}^p, b \in \mathbb{R} \tag{H.2}$$

Dans les travaux de Cortes et Vapnik [1995] sont introduites les marges souples et ainsi les variables $\xi_i \geq 0$ mesurant le degré d'erreur de classification. Par conséquent, les contraintes deviennent :

$$w.x_i + b \geq 1 - \xi_i \quad \text{si} \quad y_i = +1 \tag{H.3}$$

$$w.x_i + b \leq 1 - \xi_i \quad \text{si} \quad y_i = -1 \tag{H.4}$$

Cela revient alors à résoudre le problème :

$$\text{argmin} \left\{ \frac{\|w\|^2}{2} + C \sum_{i}^{n} \xi_i \right\} \text{ sous contrainte } \; y_i (w.x_i + b) \geq 1 - \xi_i \tag{H.5}$$

avec C le facteur de coût contrôlant le compromis entre les erreurs de classification (*outliers)* et la marge. Il est possible d'utiliser différents facteurs de coût C_+ et C_-, respectivement associés aux classes positives et négatives, dans le cas d'ensembles d'apprentissage déséquilibrés et ainsi

éviter que la solution ne soit biaisée par la sur-représentation d'une classe par rapport à l'autre (voir Morik *et al.* [1999]).

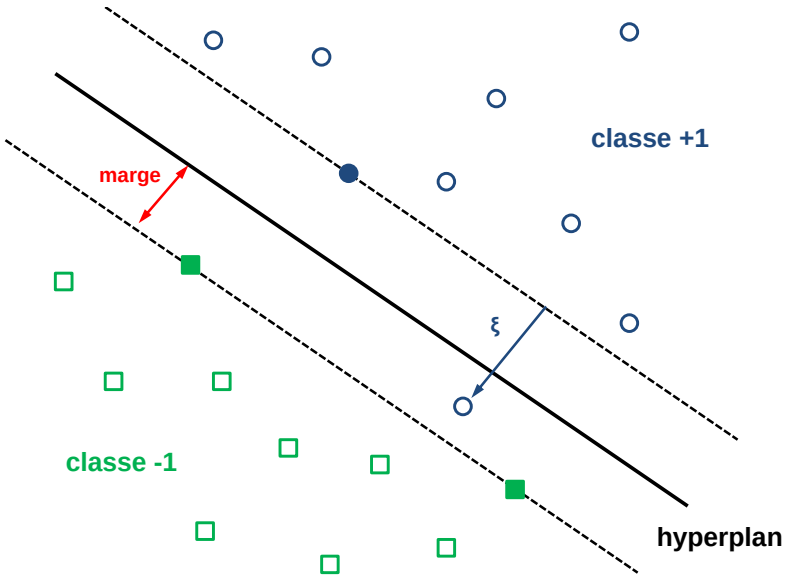


Figure H.1 – Classification SVM avec maximisation de la marge séparatrice. Les vecteurs de support sont les points sur lesquels s'appuient les marges.

L'astuce du noyau (*kernel trick*) permet de projeter les observations dans un espace de Hilbert $\mathscr{E}$ associé avec un noyau $k(\cdot,\cdot)$ et la transformation $\Phi(x) = k(x,\cdot)$ (RKHS).

Un vecteur de la base de test est alors classé suivant le signe de la fonction :

$$g(x) = \sum_{i=1}^{n_s} \alpha_i y_i k(s_i, x) + b \tag{H.6}$$

avec s_i les vecteurs de support, α_i les multiplicateurs de Lagrange, et n_s le nombre de vecteurs de support. Par conséquent la classification linéaire dans l'espace transformé est l'équivalent d'une classification non-linéaire dans l'espace de départ. Plus de détails peuvent être trouvés dans les ouvrages de Cristianini et Shawe-Taylor [2000] et Schölkopf et Smola [2001].

Liste des publications

Peter WILKINS, Tomasz ADAMEK, Daragh BYRNE, Gareth J. F. JONES, Hyowon LEE, Gordon KEENAN, Kevin MCGUINESS, Noel E. O'CONNOR, Alan F. SMEATON, Alia AMIN, Zeljko OBRENOVIC, Rachid BENMOKHTAR, Eric GALMAR, Benoît HUET, Slim ESSID, Rémi LANDAIS, Félicien VALLET, Georgios Th. PAPADOPOULOS, Stefanos VROCHIDIS, Vasileios MEZARIS, Ioannis KOMPATSIARIS, Evaggelos SPYROU, Yannis AVRITHIS, Roland MORZINGER, Peter SCHALLAUER, Werner BAILER, Tomas PIATRIK, Krishna CHANDRAMOULI, Ebroul IZQUIERDO,Martin HALLER, Lutz GOLDMANN, Amjad SAMOUR, Andreas COBET, Thomas SIKORA et Pavel PRAKS : K-Space at TRECVid 2007. *Dans les actes de : TRECVID 2007 Workshop*, Gaithersburg,MD, USA, novembre 2007.

Félicien VALLET, Gaël RICHARD, Slim ESSID, Jean CARRIVE : Detecting artist performances in a TV show. *Dans les actes de : K-Space PhD Jamboree*, Paris, France, juillet 2008.

Zaïd HARCHAOUI, Félicien VALLET, Alexandre LUNG-YUT-FONG et Olivier CAPPÉ : A regularized kernel-based approach to unsupervised audio segmentation. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, avril 2009.

Félicien VALLET, Slim ESSID, Jean CARRIVE, Gaël RICHARD : Descripteurs visuels robustes pour l'identification de locuteurs dans des émissions télévisées de talk-shows. *Dans les actes de : Compression et Représentation des Signaux Audiovisuels*, Lyon, France, octobre 2010.

Simon BOZONNET, Félicien VALLET, Nick EVANS, Slim ESSID, Gaël RICHARD et Jean CARRIVE : A multimodal approach to initialisation for top-down speaker diarization of television shows. *Dans les actes de : European Signal Processing Conference*, Aalborg, Denmark, août 2010.

Félicien VALLET, Slim ESSID, Jean CARRIVE et Gaël RICHARD : Robust visual features for the multimodal identification of unregistered speakers. *Dans les actes de : International Conference on Image Processing*, Hong-Kong, China, octobre 2010.

Félicien VALLET, Slim ESSID, Jean CARRIVE et Gaël RICHARD : *High-Level TV talk show structuring centered on speakers' interventions*. Chapter for TV content analysis : techniques and applications, (Yiannis Kompatsiaris, Bernard Merialdo et Shiguo Lian, éditeurs), CRC Press, Taylor Francis LLC à paraître en février 2012.

Bibliographie

Conservation à long terme des documents numérisés. Rapport technique, Ministère de la Culture et de la Communication, 2008. 40

Philippe AIGRAIN, Philippe JOLY et Véronique LONGUEVILLE : Medium-knowledge-based macrosegmentation of video into sequences. *Intelligent Multimedia Information Retrieval*, 25:74 – 84, 1997. 59

Jitendra AJMERA et Chuck WOOTERS : A robust speaker clustering algorithm. *Dans les actes de : Workshop on Automatic Speech Recognition Understanding*, St. Thomas, U.S. Virgin Islands, décembre 2003. 81

AMI : AMI Multimedia Multimodal Meetings Database - http ://www.amiproject.org/business-portal/research-approach/ami-multimedia-multimodal-meetings-database. 81

Xavier ANGUERA : *Robust speaker diarization for meetings*. Thèse de doctorat, Universitat Politècnica de Catalunya, Spain, 2006. 76, 86

Xavier ANGUERA, Simon BOZONNET, Nicholas EVANS, Corinne FREDOUILLE, Gerald FRIEDLAND et Oriol VINYALS : Speaker diarization : a review of recent research. *IEEE Transactions On Acoustics Speech and Language Processing*, to be published, 2011. 76

Xavier ANGUERA, Chuck WOOTERS et Jose M. PARDO : Robust speaker diarization for meetings : ICSI RT06s meetings evaluation system. *Lecture Notes in Computer Science*, 4299(3):346 – 358, 2006. 80

Hisashi AOKI, Shigeyoshi SHIMOTSUJI et Osamu HORI : A shot classification method of selecting effective key-frames for video browsing. *Dans les actes de : ACM International Conference on Multimedia*, Boston, MA, USA, novembre 1996. 59

José Anibal ARIAS, Julien PINQUIER et Régine ANDRÉ-OBRECHT : Evaluation of classification techniques for audio indexing. *Dans les actes de : European Signal Processing Conference*, Antalya, Turkey, septembre 2005. 34, 46

Jürgen ASSFALG, Alberto del BIMBO, Walter NUNZIATI et Pietro PALA : Soccer highlights detection and recognition using HMMs. *Dans les actes de : International Conference on Multimedia and Expo*, Lausanne, Switzerland, août 2002. 16, 62

Olivier AUBERT et Yannick PRIÉ : Advene : an open-source framework for integrating and visualising audiovisual metadata. *Dans les actes de : ACM International Conference on Multimedia*, Augsburg, Germany, septembre 2007. 97

Tom BACKSTROM et Carlo MAGI : Properties of line spectrum pair polynomials - a review. *Signal Processing*, 86(11):3286 – 3298, 2006. 113

Siwar BAGHDADI, Guillaume GRAVIER, Claire-Hélène DEMARTY et Patrick GROS : Structure learning in bayesian network based video indexing. *Dans les actes de : International Conference on Multimedia and Expo*, Hannover, Germany, juin 2008. 56, 62

Mark BAILLIE et Joemon M. JOSE : An audio-based sports video segmentation and event detection algorithm. *Dans les actes de : Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, USA, juin 2004. 16, 62

Claude BARRAS, Xuan ZHU, Sylvain MEIGNIER et Jean-Luc GAUVAIN : Multistage speaker diarization of broadcast news. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 14 (5):1505 – 1512, 2006. 80

Mathieu BEN, Michaël BETSER, Frédéric BIMBOT et Guillaume GRAVIER : Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. *Dans les actes de : International Conference on Spoken Language Processing*, Jeju Island, Korea, octobre 2004. 80, 81

Meriem BENDRIS : *Indexation audio-visuelle des personnes dans un contexte de télévision.* Thèse de doctorat, TELECOM ParisTech, France, 2011. 17, 32, 34, 141, 146

Meriem BENDRIS, Delphine CHARLET et Gérard CHOLLET : Lip activity detection for talking faces classification in TV-content. *Dans les actes de : International Conference on Machine Vision*, Hong-Kong, China, décembre 2010a. 83

Meriem BENDRIS, Delphine CHARLET et Gérard CHOLLET : Talking faces indexing in TV-content. *Dans les actes de : Content-Based Multimedia Indexing*, Grenoble, France, juin 2010b. 34, 83

Rachid BENMOKHTAR, Eric GALMAR et Benoit HUET : EURECOM at TRECVid 2007 : extraction of high level features. *Dans les actes de : International Workshop on Video Retrieval Evaluation*, Gaithersburg, MD, USA, novembre 2007. 62

Benjamin BIGOT, Isabelle FERRANÉ et Julien PINQUIER : Exploiting speaker segmentations for automatic role detection. An application to broadcast news documents. *Dans les actes de : International Workshop on Content-Based Multimedia Indexing*, Grenoble, France, juin 2010a. 141

Benjamin BIGOT, Isabelle FERRANÉ, Julien PINQUIER et Régine ANDRÉ-OBRECHT : Speaker role recognition to help spontaneous conversational speech detection. *Dans les actes de : ACM Workshop on Searching for Spontaneous Conversational Speech*, Firenze, Italy, octobre 2010b. 141

Isabelle BLOCH : *Fusion d'informations en traitement du signal et des images*. Hermes Lavoisier, 2003. 63

Jean-François BONASTRE : La reconnaissance du locuteur : un problème résolu ? *Dans les actes de : Journées d'études sur la Parole*, Avignon, France, juin 2008. 86

Pierre BOURDIEU : *Sur la télévision*. Raisons d'agir, 1996. 30

Jérôme BOURDON : Propositions pour une semiologie des genres audiovisuels. *Quaderni*, 4:19 – 36, 1988. 27, 28

Simon BOZONNET, Nicholas EVANS et Corinne FREDOUILLE : The LIA-EURECOM RT'09 speaker diarization system : enhancements in speaker modelling and cluster purification. *Dans les actes de : International Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, mars 2010a. 10, 87, 88, 89, 102, 104, 108

Simon BOZONNET, Félicien VALLET, Nick EVANS, Slim ESSID, Gael RICHARD et Jean CARRIVE : A multimodal approach to initialisation for top-down speaker diarization of television shows. *Dans les actes de : European Signal Processing Conference*, Aalborg, Denmark, août 2010b. 10, 34, 46, 64, 85

Gary BRADSKI et Adrian KAEHLER : *Learning OpenCV : computer vision with the OpenCV library*. O'Reilly Media, 2008. 95, 169

Hervé BREDIN et Gérard CHOLLET : Measuring audio and visual speech synchrony : methods and applications. *Dans les actes de : International Conference on Visual Information Engineering*, Bangalore, India, septembre 2006. 82

Hervé BREDIN et Gérard CHOLLET : Audio-visual speech synchrony measure for talking-face identity verification. *Dans les actes de : International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, avril 2007. 83, 125

Paul BROWNE, Alan SMEATON, Noel MURPHY, Noel E. O'CONNOR, Sean MARLOW et Catherine BERRUT : Evaluation and combining digital video shot boundary detection algorithms. *Dans les actes de : Irish Machine Vision and Information Processing Conference*, Belfast, United Kingdom, septembre 2000. 59

Roberto BRUNELLI, Ornella MICH et Carla-Maria MODENA : A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78 – 112, 1999. 13, 59

Marine CAMPEDEL et Pierre HOOGSTOËL, éditeurs. *Sémantique et multimodalité en analyse de l'information*. Hermes Science Publications, 2011. 57, 63

CANAL9 : Canal 9 political debates database - http ://canal9-db.sspnet.eu/. 81

Jean CARLETTA, Simone ASHBY, Sebastien BOURBAN, Mike FLYNN, Mael GUILLEMOT, Thomas HAIN, Jaroslav KADLEC, Vasilis KARAISKOS, Wessel KRAAIJ, Melissa KRONENTHAL, Guillaume LATHOUD, Mike LINCOLN, Agnes LISOWSKA, Iain MCCOWAN, Wilfried POST, Dennis REIDSMA

et Pierre WELLNER : The AMI meeting corpus : a pre-announcement. *Dans les actes de : Machine Learning for Multimodal Interaction : Second International Workshop*, Edinburgh , United Kingdom, juillet 2005. 81

Jean CARRIVE : Document description for audiovisual archiving, corpora, technologies and uses. *Dans les actes de : Content-Based Multimedia Indexing*, Bordeaux, France, juin 2007. 15

Sabine CHALVON-DEMERSAY et Dominique PASQUIER : Le langage des variétés. *Terrain*, 15:29 – 40, 1990. 30, 37, 43

Chih-Chung CHANG et Chih-Jen LIN : LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27 :1 – 27 :27, 2011. 119

Patrick CHARAUDEAU : Les conditions d'une typologie des genres télévisuels d'information. *Réseaux*, 81:79 – 101, 1997. 27, 28, 30

Scott S. CHEN et P.S. GOPALAKRISHNAN : Speaker, environment and channel change detection and clustering via the bayesian information criterion. *Dans les actes de : DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, février 1998. 59, 79

Corinna CORTES et Vladimir VAPNIK : Support-vector networks. *Machine Learning*, 20(3):273 – 297, 1995. 171

Timothée COUR, Chris JORDAN, Eleni MILTSAKAKI et Ben TASKAR : Movie/script : alignment and parsing of video and text transcription. *Dans les actes de : European Conference on Computer Vision*, Marseille, France, octobre 2008. 16

Nello CRISTIANINI et John SHAWE-TAYLOR : *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000. 60, 172

Perrine DELACOURT et Christian J. WELLEKENS : DISTBIC : a speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1 - 2):111 – 126, 2000. 80

Manolis DELAKIS : *Multimodal tennis video structure analysis with segment models*. Thèse de doctorat, Université de Rennes, France, 2006. 16

Manolis DELAKIS, Guillaume GRAVIER et Patrick GROS : Audiovisual integration with segment models for tennis video parsing. *Computer Vision and Image Understanding*, 111(2):142 – 154, 2008. 16, 62

Paul DELÉGLISE, Yannick ESTÈVE, Sylvain MEIGNIER et Teva MERLIN : The LIUM speech transcription system : a CMU Sphinx iii-based system for french broadcast news. *Dans les actes de : International Speech Communication Association*, Lisbon, Portugal, septembre 2005. 17

Alfred DIELMANN : Unsupervised detection of multimodal clusters in edited recordings. *Dans les actes de : Multimedia Signal Processing*, Saint-Malo, France, octobre 2010. 16, 83

Nevenka DIMITROVA, Hong-Jiang ZHANG, Behzad SHAHRARAY, Ibrahim SEZAN, Thomas HUANG et Avideh ZAKHOR : Applications of video-content analysis and retrieval. *IEEE Transactions on Multimedia*, 9(3):42 – 55, 2002. 13

Frédéric DÉSOBRY, Manuel DAVY et Christian DONCARLI : An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961 – 2974, 2005. 79

Richard O. DUDA, Peter E. HART et David G. STORK : *Pattern classification*. Wiley, 2000. 59, 77, 100, 103

Slim ESSID : *Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique*. Thèse de doctorat, TELECOM ParisTech, France, 2005. 118

ESTER : Campagne d'évaluation des systèmes de transcription enrichie d'emissions radiophoniques - http ://www.afcp-parole.org/camp_eval_systems_transcriptions/. 76

Mark EVERINGHAM, Josef SIVIC et Andrew ZISSERMAN : Hello ! My name is... Buffy – automatic naming of characters in TV video. *Dans les actes de : British Machine Vision Conference*, Edinburgh, United Kingdom, septembre 2006. 114, 115

Pedro F. FELZENSZWALB et Daniel P. HUTTENLOCHER : Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):57 – 79, 2005. 114

Belkacem FERGANI, Manuel DAVY et Amrane HOUACINEA : Speaker diarization using one-class support vector machines. *Speech Communication*, 50(5):355 – 365, 2008. 80, 81

John FISHER et Trevor DARRELL : Signal level fusion for multimodal perceptual user interface. *Dans les actes de : Workshop on Perceptive user interfaces*, Orlando, FL, USA, novembre 2001. 83

John FISHER, Trevor DARRELL, William FREEMAN et Paul VIOLA : Learning joint statistical models for audio-visual fusion and segregation. *Dans les actes de : Neural Information Processing Systems*, Vancouver, BC, Canada, décembre 2000. 83

Corinne FREDOUILLE, Simon BOZONNET et Nicholas EVANS : The LIA-EURECOM RT'09 speaker diarization system. *Dans les actes de : RT'09 NIST Rich Transcription Workshop*, Melbourne, FL, USA, mai 2009. 10, 56, 78, 81, 87, 88, 89, 93, 102, 104, 108

Yoav FREUND et Robert E. SCHAPIRE : A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. 96

Yoav FREUND et Robert E. SCHAPIRE : A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771 – 780, 1999. 96

Gerald FRIEDLAND, Hayley HUNG et Chuohao YEO : Multi-modal speaker diarization of real-world meetings using compressed-domain video features. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, avril 2009a. 81, 83, 93, 99

Gerald FRIEDLAND, Oriol VINYALS, Yan HUANG et Christian MÜLLER : Prosodic and other long-term features for speaker diarization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 17(5):985 – 993, 2009b. 76, 81

Gerald FRIEDLAND, Chuohao YEO et Hayley HUNG : Visual speaker localization aided by acoustic models. *Dans les actes de : ACM International Conference on Multimedia*, Beijing, China, avril 2009c. 83, 108

Sylvain GALLIANO, Guillaume GRAVIER et Laura CHAUBARD : The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *Dans les actes de : Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, septembre 2009. 76

Rodolphe GHIGLIONE et Patrick CHARAUDEAU : *La Parole confisquée. Un genre télévisuel : le talk show*. Dunod, 1997. 30

Herbert GISH, Man-Hung SIU et Robin ROHLICEK : Segregation of speakers for speech recognition and speaker identification. *Dans les actes de : International Conference on Acoustics, Speech, and Signal Processing*, Toronto, ON, Canada, mai 1991. 59, 80

Erving GOFFMAN : *Forms of talk*. University of Pennsylvania Press, 1981. 30

Yihong GONG, Lim SIN et Chua CHUAN : Automatic parsing of TV soccer programs. *Dans les actes de : International Conference on Multimedia Computing and Systems*, Washington, DC, USA, mai 1995. 62

Camille GUINAUDEAU, Guillaume GRAVIER et Pascale SÉBILLOT : Can automatic speech transcripts be used for large scale TV stream description and structuring ? *Dans les actes de : Workshop on Content-Based Audio Video Analysis for Novel TV Services*, San Diego, CA, USA, december 2009. 17

Isabelle GUYON et André ELISSEEFF : An introduction to feature and variable selection. *Journal of Machine Learning Research*, 3:1157 – 1182, 2003. 100

André GUÉZIEC : Tracking pitches for broadcast television. *IEEE Computer*, 35(3):38 – 43, 2002. 16, 62

Raffay HAMID, Ram Krishan KUMAR, Matthias GRUNDMANN, Kihwan KIM, Irfan ESSA et Jessica HODGINS : Player localization using multiple static cameras for sports visualization. *Dans les actes de : Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, juin 2010. 62

Yina HAN, Guizhong LIU, Gérard CHOLLET et Joseph RAZIK : Person identity clustering in TV show videos. *Dans les actes de : Visual Information Engineering*, Xian, China, juillet 2008. 34, 46

Zaïd HARCHAOUI, Félicien VALLET, Alexandre LUNG-YUT-FONG et Olivier CAPPÉ : A regularized kernel-based approach to unsupervised audio segmentation. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, avril 2009. 34, 46, 59, 80

Chris HARRIS et Mike STEPHENS : A combined corner and edge detector. *Dans les actes de : Alvey Vision Conference*, Manchester, United Kingdom, septembre 1988. 168, 169

John HERSHEY et Javier MOVELLAN : Audio-vision : Using audiovisual synchrony to locate sounds. *Advances in Neural Information Processing (MIT Press)*, pages 813 – 819, 2000. 82

Cyril HORY et William J. CHRISTMAS : Cepstral features for classification of an impulse response with varying sample size dataset. *Dans les actes de : European Signal Processing Conference*, Poznan, Poland, septembre 2007. 56, 61

Harold HOTELLING : Relations between two sets of variates. *Biometrika*, 28(3 - 4):321 – 377, 1936. 124

Hayley HUNG et Gerald FRIEDLAND : Towards audio-visual on-line diarization of participants in group meetings. *Dans les actes de : Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, Marseille, France, octobre 2008. 83

Ichiro IDE, Hiroshi MO, Norio KATAYAMA et Shinichi SATOH : Topic threading for structuring a large-scale news video archive. *Lecture Notes in Computer Science*, 3115(1):2128–2129, 2004. 16

Gaël JAFFRÉ : *Indexation de la vidéo par le costume*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 2005. 16, 95

Gaël JAFFRÉ et Philippe JOLY : Costume : a new feature for automatic video content indexing. *Dans les actes de : International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*, Avignon, France, avril 2004. 95

Elie El KHOURY : *Unsupervised video indexing based on audiovisual characterization of persons*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, 2010. 17

Elie El KHOURY, Sylvain MEIGNER et Christine SÉNAC : Speaker diarization : combination of the LIUM and IRIT systems. Rapport technique, IRIT / LIUM, 2008. 80, 81

Elie El KHOURY, Christine SENAC et Philippe JOLY : Face-and-clothing based people clustering in video content. *Dans les actes de : ACM International Conference on Multimedia Information Retrieval*, Philadelphia, PA, USA, mars 2010. 95

Ewa KIJAK, Guillaume GRAVIER, Lionel OISEL et Patrick GROS : Audiovisual integration for tennis broadcast structuring. *Multimedia Tools and Applications*, 30(3):289 – 311, 2006. 16, 62

Kihwan KIM, Matthias GRUNDMANN, Ariel SHAMIR, Iain MATTHEWS, Jessica HODGINS et Irfan ESSA : Motion fields to predict play evolution in dynamic sport scenes. *Dans les actes de : Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, juin 2010. 62

Julien LAW-TO, Gregory GREFENSTETE, Jean-Luc GAUVAIN, Guillaume GRAVIER, Lori LAMEL et Julien DESPRES : VoxaleadNews : robust automatic segmentation of video content into browsable and searchable subjects. *Dans les actes de : ACM Multimedia*, Firenze, Italy, octobre 2010. 16

Marie LHÉRAULT et Erik NEVEU : Quelques dispositifs de talk-shows français (1998-2003). *Réseaux*, 118:201 – 207, 2003. 30

Dongge LI, Nevenka DIMITROVA, Mingkun LI et Ishwar SETHI : Multimedia content processing through cross-modal association. *Dans les actes de : ACM International Conference on Multimedia*, Berkeley, CA, USA, novembre 2003. 83, 125

Rainer LIENHART : Comparison of automatic shot boundary detection algorithms. *Dans les actes de : Storage and retrieval for image and video databases*, San Jose, CA, USA, décembre 1998. 59

Guy LOCHARD : Débats, talk-shows : de la radio filmée ? *Communication et langages*, 86(4):92 – 100, 1990. 28, 30

David G. LOWE : Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91 – 110, 2004. 56

Bruce LUCAS et Takeo KANADE : An iterative image registration technique with an application to stereo vision. *Dans les actes de : International Joint Conference on Artificial Intelligence*, Vancouver, BC, Canada, août 1981. 97, 168

Subhransu MAJI, Alexander C. BERG et Jitendra MALIK : Classification using intersection kernel support vector machines is efficient. *Dans les actes de : Computer Vision and Pattern Recognition*, Anchorage, AK, USA, juin 2008. 124

B.S. MANJUNATH, Philippe SALEMBIER et Thomas SIKORA, éditeurs. *Introduction to MPEG-7 - multimedia content description interface*. Wiley, 2002. 98, 100

Gaël MANSON et Sid-Ahmed BERRANI : Automatic TV broadcast structuring. *Journal of Digital Multimedia Broadcasting*, 2010(1):1 – 16, 2010. 16

Benoît MATHIEU, Slim ESSID, Thomas FILLON, Jacques PRADO et Gaël RICHARD : YAAFE, an easy to use and efficient audio feature extraction software. *Dans les actes de : International Society for Music Information Retrieval Conference*, Utrecht, the Netherlands, août 2010. 94

MBGC : Multiple Biometric Grand Challenge - http ://www.nist.gov/itl/iad/ig/mbgc.cfm. 82

Sylvain MEIGNIER, Jean-François BONASTRE et Stéphane IGOUNET : E-HMM approach for learning and adapting sound models for speaker indexing. *Dans les actes de : Odyssey Speaker and Language Recognition Workshop*, Crete, Greece, juin 2001. 78, 81, 88

Daniel MORARU, Mathieu BEN et Guillaume GRAVIER : Experiments on speaker tracking and segmentation in radio broadcast news. *Dans les actes de : International Conference on Spoken Language Processing*, Lisbon, Portugal, septembre 2005. 81

Katharina MORIK, Peter BROCKHAUSEN et Thorsten JOACHIMS : Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. *Dans les actes de : International Conference on Machine Learning*, Bled, Slovenia, juin 1999. 172

Wayne MUNSON : *All talk : the talk show in media culture*. Temple University Press, 1993. 30

Kevin MURPHY : *Dynamic bayesian networks : representation, inference and learning*. Thèse de doctorat, UC Berkeley, Computer Science Division, CA, USA, 2002. 62

MVGL-AVD : Multimedia, Vision and Graphics Laboratory Audio-Visual Database - http ://mvgl.ku.edu.tr/databases. 82

Xavier NATUREL et Patrick GROS : Detecting repeats for video structuring. *Multimedia Tools and Applications*, 38(2):233 – 252, 2008. 16

NIST : National Institute of Standards and Technology - http ://www.nist.gov/srd/index.htm. 75

Harriet J. NOCK, Giridharan IYENGAR et Chalapathy NETI : Speaker localisation using audio-visual synchrony : an empirical study. *Dans les actes de : International conference on image and video retrieval*, Urbana-Champaign, IL, USA, juillet 2003. 82

Nuria OLIVER et Eric HORVITZ : A comparison of HMMs and dynamic bayesian networks for recognizing office activities. *Dans les actes de : International conference on user modeling*, Edinburgh, United Kingdom, juillet 2005. 16

Eric K. PATTERSON, Sabri GURBUZ, Zekeriya TUFEKCI et John N. GOWDY : CUAVE : a new audio-visual database for multimodal human-computer interface research. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, USA, mai 2002. 82

Hermine PENZ : *Language and control in American TV talk shows : an Analysis of linguistic strategies*. Gunter Narr, 1996. 30, 37, 43

Christian PETERSOHN : Fraunhofer Heinrich Hertz Institute at TRECVID 2004 : shot boundary detection system. *Dans les actes de : TRECVID 2004 Workshop*, Gaithersburg, MD, USA, novembre 2004. 59

Milan PETKOVIC, Vojkan MIHAJLOVIC, Willem JONKER et S. DJORDJEVIC-KAJAN : Multi-modal extraction of highlights from TV formula 1 programs. *Dans les actes de : International Conference on Multimedia and Expo*, Lausanne, Switzerland, août 2002. 16, 62

Julien PINQUIER et Régine ANDRÉ-OBRECHT : Jingle detection and identification in audio documents. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Montreal, QC, Canada, mai 2004. 56, 61

John C. PLATT : *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*. The MIT Press, 1999. 120

Jean-Philippe POLI : *Structuration automatique de flux télévisuels*. Thèse de doctorat, Université Paul Cézanne, Aix-Marseille, France, 2007. 6, 58

Jean-Philippe POLI : An automatic television stream structuring system for television archives holder. *Multimedia systems*, 14(5):255 – 275, 2008. 16

Jacques PRADO : Conversion de fréquence. Rapport technique, TELECOM ParisTech, 2009. 145

Lawrence RABINER et Biing-Hwang JUANG : *Fundamentals of speech recognition*. Prentice Hall PTR, 1993. 76, 94

Lawrence R. RABINER : A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 286, 1989. 60, 77

Javier RAMÍREZ, Juan Manuel GÓRRIZ et José Carlos SEGURA : Voice activity detection. fundamentals and speech recognition system robustness. *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel:1 – 22, 2007. 78

Douglas A. REYNOLDS et Pedro A. TORRES-CARRASQUILLO : Approaches and applications of audio diarization. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA, USA, mars 2005. 59, 77

Gaël RICHARD, Mathieu RAMONA et Slim ESSID : Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, avril 2007. 56, 60, 79

Jamal-Eddine ROUGUI, Mohammed RZIZA, Driss ABOUTAJDINE, Marc GELGON et José MARTINEZ : Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, mai 2006. 77, 81

Mehmet Emre SARGIN, Yücel YEMEZ, Engin ERZIN et A. Murat TEKALP : Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Transactions on Multimedia*, 9 (7):1396 – 1403, 2007. 82

John SAUNDERS : Real-time discrimination of broadcast speech/music. *Dans les actes de : International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, USA, mai 1996. 60, 78

Eric SCHEIRER et Malcolm SLANEY : Construction and evaluation of a robust multifeature speech/music discriminator. *Dans les actes de : International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, avril 1997. 60, 78

Bernhard SCHÖLKOPF, John C. PLATT, John SHAWE-TAYLOR, Alex J. SMOLA et Robert C. WILLIAMSON : Estimating the support of a high-dimensional distribution. *Neural Computation*, 13 (7):1443 – 1471, 2001. 126, 128

Bernhard SCHÖLKOPF et Alexander J. SMOLA : *Learning with kernels : support vector machines, regularization, optimization and beyond*. MIT-Press, 2001. 128, 172

Hans W. SCHUSSLER : A stability theorem for discrete systems. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(1):87–89, 1976. 113

Jianbo SHI et Carlo TOMASI : Good features to track. *Dans les actes de : International Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, juin 1994. 97, 168, 169

Matthew A. SIEGLER, Uday JAIN, Bhiksha RAJ et Richard M. STERN : Automatic segmentation, classification and clustering of broadcast news audio. *Dans les actes de : DARPA Speech Recognition Workshop*, Chantilly, VA, USA, février 1997. 80

Josef SIVIC, Mark EVERINGHAM et Andrew ZISSERMAN : Who are you ? : Learning person specific classifiers from video. *Dans les actes de : International Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, juin 2009. 16, 114

Malcolm SLANEY et Michele COVELL : FaceSync : a linear operator for measuring synchronization of video facial images and audio tracks. *Dans les actes de : Neural Information Processing Systems*, Denver, CO, USA, novembre 2000. 82

Han SLÖETJES et Peter WITTENBURG : Annotation by category - ELAN and ISO DCR. *Dans les actes de : International Conference on Language Ressources and Evaluation*, Marrakech, Morocco, mai 2008. 32, 48

Alan SMEATON, Paul OVER et Aiden DOHERTY : Video shot boundary detection : Seven years of TRECVid activity. *Computer Vision and Image Understanding*, 114(4):1 – 25, 2009. 58

George W. SNEDECOR et William G. COCHRAN : *Statistical methods*. Iowa State University Press, 1967. 116

Cees G. M. SNOEK et Marcel WORRING : Multimodal video indexing : a review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5 – 35, 2005. 13

Catherine A. SUGAR et Gareth M. JAMES : Finding the number of clusters in a data set : an information theoretic approach. *Journal of the American Statistical Association*, 98:397–408, 2003. 103

Michael J. SWAIN et Dana H. BALLARD : Color indexing. *International Journal of Computer Vision*, 7(1):11 – 32, 1991. 56, 124

Bernard TIMBERG : *Television talk : a history of TV talk show*. University of Texas Press, 2002. 30

Sue E. TRANTER et Douglas A. REYNOLDS : An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557 – 1565, 2006. 76

Wei-Ho TSAI, Shih-Sian CHENG et Hsin-Min WANG : Speaker clustering of speech utterances using a voice characteristic reference space. *Dans les actes de : International Conference on Spoken Language Processing*, Jeju Island, Korea, octobre 2004. 77

Himanshu VAJARIA, Tanmoy ISLAM, Sudeep SARKAR, Ravi SANKAR et Ranga KASTURI : Audio segmentation and speaker localization in meeting videos. *Dans les actes de : International Conference on Pattern Recognition*, Hong Kong, China, août 2006. 83

Félicien VALLET, Slim ESSID, Jean CARRIVE et Gaël RICHARD : Robust visual features for the multimodal identification of unregistered speakers. *Dans les actes de : International Conference on Image Processing*, Hong-Kong, China, octobre 2010. 34, 46, 86, 108, 111, 112, 119, 140

Félicien VALLET, Gaël RICHARD, Slim ESSID et Jean CARRIVE : Detecting artist performances in a TV show. *Dans les actes de : K-Space PhD Jamboree*, Paris, France, juillet 2008. 141

Jeroen VENDRIG et Marcel WORRING : Systematic evaluation of logical story unit segmentation. *IEEE Transactions on Multimedia*, 4(4):492 – 499, 2002. 59

Alessandro VINCIARELLI, Alfred DIELMANN, Sarah FAVRE et Hugues SALAMIN : Canal9 : a database of political debates for analysis of social interactions. *Dans les actes de : International Conference on Affective Computing and Intelligent Interaction*, Amsterdam, the Netherlands, septembre 2009. 81

Paul VIOLA et Michael JONES : Robust real-time object detection. *Dans les actes de : International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling*, Vancouver, BC, Canada, juillet 2001. 95, 114

Timo VOLKMER, Seyed TAHAGHOGHI et James THOM : RMIT university video retrieval experiments at TRECVid 2004. *Dans les actes de : TRECVID 2004 Workshop*, Gaithersburg, MD, USA, novembre 2004. 59

Howard D. WACTLAR, Takeo KANADE, Michael A. SMITH et Scott M. STEVENS : Intelligent access to digital video : Informedia project. *IEEE Computer*, 29(5):46 – 52, 1996. 59

Peter WILKINS, Tomasz ADAMEK, Daragh BYRNE, Gareth J. F. JONES, Hyowon LEE, Gordon KEENAN, Kevin MCGUINESS, Noel E. O'CONNOR, Alan F. SMEATON, Alia AMIN, Zeljko OBRENOVIC, Rachid BENMOKHTAR, Eric GALMAR, Benoît HUET, Slim ESSID, Rémi LANDAIS, Félicien VALLET, Georgios Th. PAPADOPOULOS, Stefanos VROCHIDIS, Vasileios MEZARIS, Ioannis KOMPATSIARIS, Evaggelos SPYROU, Yannis AVRITHIS, Roland MORZINGER, Peter SCHALLAUER, Werner BAILER, Tomas PIATRIK, Krishna CHANDRAMOULI, Ebroul IZQUIERDO, Martin HALLER, Lutz GOLDMANN, Amjad SAMOUR, Andreas COBET, Thomas SIKORA et Pavel PRAKS : K-Space at TRECVid 2007. *Dans les actes de : TRECVID 2007 Workshop*, Gaithersburg, MD, USA, novembre 2007. 56, 61

Gethin WILLIAMS et Daniel P.W. ELLIS : Speech/music discrimination based on posterior probability features. *Dans les actes de : European Conference on Speech Communication and Technology*, Budapest, Hungary, septembre 1999. 78

Raymond WILLIAMS : *Television : technology and cultural form*. Fontana, 1974. 28

Chuck WOOTERS, James FUNG, Barbara PESKIN et Xavier ANGUERA : Towards robust speaker segmentation : the ICSI-SRI fall 2004 diarization system. *Dans les actes de : Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, USA, novembre 2004. 79

Chuck WOOTERS et Marijn HUIJBREGTS : The ICSI RT'07s speaker diarization system. *Multimodal Technologies for Perception of Humans*, 4625:509 – 519, 2008. 77, 79, 81, 83

Ting-Fan WU, Chih-Jen LIN et Ruby C. WENG : Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975 – 1005, 2004. 120

Ziyou XIONG, Regunathan RADHAKRISHNAN et Ajay DIVAKARAN : Generation of sports highlights using motion activities in combination with a common audio feature extraction framework. *Dans les actes de : International Conference on Image Processing*, Barcelona, Spain, septembre 2003. 16, 62

XM2VTSDB : The Extended Multi Modal Verification for Teleservices and Security applications DataBase - http ://www.ee.surrey.ac.uk/cvssp/xm2vtsdb/. 82

Min XU, Ling-Yu DUAN, Changsheng XU, Mohan KANKANHALLI et Qi TIAN : Event detection in basketball video using multiple modalities. *Dans les actes de : Pacific-Rim Conference on Multimedia*, Singapore, décembre 2003. 16, 62

Minerva YEUNG et Boon-Lock YEO : Time-constrained clustering for segmentation of video into story units. *Dans les actes de : International Conference on Pattern Recognition*, Vienna, Austria, août 1996. 59

Xinguo YU, Liyuan LI et Hon Wai LEONG : Interactive broadcast services for live soccer video based on instant semantics acquisition. *Visual Communication and Image Representation*, 20 (2):117 – 130, 2009. 16, 62

Dongqing ZHANG et Shih-Fu CHANG : Event detection in baseball video using superimposed caption recognition. *Dans les actes de : ACM Conference on Multimedia*, Juan les Pins, France, décembre 2002. 16, 62

Shaohua Kevin ZHOU et Rama CHELLAPPA : From sample similarity to ensemble similarity : probabilistic distance measures in reproducing kernel hilbert space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):917 – 929, 2006. 118

Wensheng ZHOU, Asha VELLAIKAL et C.-C. J. KUO : Rule based video classification system for basketball video indexing. *Dans les actes de : ACM workshops on Multimedia*, Los Angeles, CA, USA, novembre 2000. 16, 62


AH PATRON! C'QUE VOUS ÊTES FORT!
BAH-BAH-BAH! C'EST L'MÉTIER MON P'TIT, C'EST L'MÉTIER!

Structuration automatique de talk shows télévisés

Les problématiques modernes de conservation du patrimoine numérique ont rendu les compagnies professionnelles d'archivage demandeuses de nouveaux outils d'indexation et en particulier de méthodes de structuration automatique. Dans cette thèse, nous nous intéressons à un genre télévisuel à notre connaissance peu analysé : le talk show.

Inspirés de travaux issus de la communauté des sciences humaines et plus spécifiquement d'études sémiologiques, nous proposons, tout d'abord, une réflexion sur la structuration d'émissions de talk show. Ensuite, ayant souligné qu'un schéma de structuration ne peut avoir de sens que s'il s'inscrit dans une démarche de résolution de cas d'usage, nous proposons une évaluation de l'organisation ainsi dégagée au moyen d'une expérience utilisateur. Cette dernière met en avant l'importance des locuteurs et l'avantage d'utiliser le tour de parole comme entité atomique en lieu et place du plan (*shot*), traditionnellement adopté dans les travaux de structuration.

Ayant souligné l'importance de la segmentation en locuteurs pour la structuration d'émissions de talk show, nous y consacrons spécifiquement la seconde partie de cette thèse. Nous proposons tout d'abord un état de l'art des techniques utilisées dans ce domaine de recherche et en particulier des méthodes non-supervisées. Ensuite sont présentés les résultats d'un premier travail de détection et regroupement des tours de parole. Puis, un système original exploitant de manière plus efficace l'information visuelle est enfin proposé. La validité de la méthode présentée est testée sur les corpus d'émissions *Le Grand Échiquier* et *On n'a pas tout dit*. Au regard des résultats, notre dernier système se démarque avantageusement des travaux de l'état de l'art. Il conforte l'idée que les caractéristiques visuelles peuvent être d'un grand intérêt — même pour la résolution de tâches supposément exclusivement audio comme la segmentation en locuteurs — et que l'utilisation de méthodes à noyau dans un contexte multimodal peut s'avérer très performante.

Mots-clés : talk show, structuration, segmentation en locuteurs, multimodalité, corrélation audiovisuelle, classification SVM, apprentissage non-supervisé

Automatic structuring of TV talk show programs

Archives professionals have high expectations for efficient indexing tools. In particular, the purpose of archiving TV broadcasts has created an expanding need for automatic content structuring methods. In this thesis, is addressed the task of structuring a particular type of TV content that has been scarcely studied in previous works, namely talk show programs.

The object of this work is examined in the light of a number of sociological studies, with the aim to identify relevant prior knowledge on the basis of which the structuring approach is motivated. Then, having highlighted that a structuring scheme should be assessed according to specific use cases, a user-based evaluation is undertaken. The latter stresses out the relevance of considering the speakers' interventions as elementary structural units instead of video shots usually employed in similar studies.

Having emphasised the importance of speaker oriented detectors, the second part of this thesis is thus put on speaker diarization methods. We first propose a state of the art of the techniques — particularly unsupervised ones — used in this research domain. Then, results on a first speaker diarization system are presented. Finally, a more original system exploiting efficiently audiovisual information is finally proposed. Its validity is tested on two talk show collections : *Le Grand Échiquier* and *On n'a pas tout dit*. The results show that this new system outperforms state of the art methods. Besides, it strengthens the interest of using visual cues — even for tasks that are considered to be exclusively audio such as speaker diarization — and kernel methods in a multimodal context.

Keywords : talk show, automatic structuring, speaker diarization, multimodality, audiovisual correlation, SVM classification, unsupervised learning