Recherche du rôle des intervenants et de leurs interactions pour la structuration de documents audiovisuels

Benjamin Bigot

THESE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'***Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)***
spécialité : ***Informatique***

Présentée et soutenue par ***Benjamin Bigot***
Le ***6 juillet 2011***

Titre :

Recherche du rôle des intervenants et de leurs interactions pour la structuration de documents audiovisuels

---

JURY

Frédéric Béchet	Professeur, Université de la Méditerranée	Président
Yannick Estève	Professeur, Université du Maine	Examinateur
Martha Larson	Senior Researcher, Delft Univ. of Technology	Rapporteur
Guillaume Gravier	Chargé de Recherche, CNRS/IRISA Rennes I	Rapporteur
Régine André-Obrecht	Professeur, Université Paul Sabatier Toulouse III	Directrice
Isabelle Ferrané	MdC, Université Paul Sabatier Toulouse III	Encadrant
Julien Pinquier	MdC, Université Paul Sabatier Toulouse III	Encadrant

---

Ecole doctorale : *Mathématiques Informatique Télécommunications (MITT)*
Unité de recherche : *Institut de Recherche en Informatique de Toulouse - UMR 5505*
Directrice de Thèse : *Pr. Régine André-Obrecht*

Remerciements

Mes remerciements vont en premier lieu à mes trois encadrants – Régine André-Obrecht, Isabelle Ferrané et Julien Pinquier – pour leur confiance et l'attention qu'ils ont porté à mon travail.

Je remercie ensuite les membres de mon jury – Martha Larson, Guillaume Gravier, Frédéric Béchet et Yannick Estève – pour leurs remarques et leurs encouragements.

Durant ces années à Toulouse, j'ai fait de nombreuses rencontres, enrichissantes professionnellement et humainement.

J'ai une pensée pour les membres de l'équipe S2I du Laboratoire d'Astrophysique de Toulouse-Tarbes, avec lesquels j'ai fait mon stage de Master Recherche.

Je remercie chaleureusement les membres de l'équipe SAMoVA parmi lesquels j'ai évolué durant ma thèse. Ils ont contribué à leur manière à l'aboutissement de mon travail.

Je remercie également l'équipe enseignante en électronique et traitement du signal de l'ENSEEIHT, avec qui j'ai eu la chance de travailler ces dernières années.

Merci à ma famille et à mes amis, qui m'ont soutenu dans cette entreprise depuis le début.

Merci Hélène d'avoir été présente, de m'avoir accompagné, supporté et toujours encouragé durant cette période.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Depuis la fin des années 80 - qui ont vu naître les premiers supports de stockage numériques ayant une capacité suffisante pour contenir à la fois du texte, du son, des images, de la vidéo - notre société vit une véritable révolution vers le « tout numérique ».

Ce mouvement émerge de la convergence :

- des innovations informatiques à travers l'augmentation des capacités de stockage, de la puissance des ordinateurs et de la miniaturisation des dispositifs électroniques,
- des avancées dans le domaine des télécommunications terrestres et satellites avec internet, la télévision numérique, la téléphonie mobile,
- de l'ouverture au public, aux professionnels et aux médias à ce type de format.

Il en résulte une explosion de la quantité de données numériques créées, archivées ou échangées chaque jour, de par le monde.

Cette évolution s'est étendue aux acteurs du « dépôt légal ». Il s'agit d'un texte inscrit au code du patrimoine ayant pour vocation la protection et la conservation du patrimoine culturel national. La France soumet au dépôt légal les auteurs de livres, périodiques, gravures, films, enregistrements sonores, émissions de radio et de télévision. Depuis la modification du texte en 2006 [1], les logiciels et les bases de données sont également concernées, dans le même temps un archivage d'une partie des sites web français devra être réalisé.

La Bibliothèque nationale de France (BnF) est responsable de la collecte et de la conservation d'une partie du dépôt légal. Son catalogue complet rassemble 14 millions de livres et d'imprimés, et depuis 1990, elle numérise toute ses nouvelles acquisitions.

Depuis 1997, Gallica [2], le portail de la bibliothèque numérique de la BnF, permet en accès libre via internet la consultation de plus d'un million de documents numériques : livres, périodiques, images, manuscrits, cartes, partitions, enregistrements audio, etc.

L'Institut national de l'audiovisuel (Ina), créé en 1975, est également un acteur important du dépôt légal. L'institut a pour mission :

- la conservation du patrimoine audiovisuel français,
- l'exploitation et la mise en valeur de ce patrimoine,
- l'accompagnement des évolutions du secteur audiovisuel.

1. de mise en application de la loi DADVSI (Droits d'Auteurs et Droits Voisins dans la Société d'Information)
2. http ://gallica.bnf.fr/

En 2010, le fonds de l'Ina compte plus de 3 millions d'heures de programmes provenant :
- de l'archivage de l'ensemble des chaînes hertziennes françaises, depuis leur création dans les années 1940 (télévision et radio),
- du dépôt légal qui rassemble, les chaînes hertziennes ainsi que les chaînes du câble et du satellite (un million d'heures de programmes sont captées en direct chaque année),
- de fonds privés (TF1, AFP, etc.).

Depuis 1999, l'Ina a débuté la numérisation de quelques 800000 heures de documents, notamment pour les préserver de la destruction de leurs supports. À la même époque plus de 100000 émissions de radio et de télévision (pour un total de 10000 heures d'archives) sont déjà consultables en ligne sur le portail web de l'Institut [3].

Les émissions de radio et de télévision archivées sont généralement décrites manuellement ou semi-manuellement, dans des notices indiquant la nature, le sujet, le producteur et toutes autres informations pouvant être utilisées pour une recherche ultérieure. Ceci représente une opération extrêmement coûteuse.

Le passage aux formats numériques a dans le même temps radicalement modifié l'attitude du public vis-à-vis des documents multimédias. Les moyens facilitant la production de photographies, de vidéos et musiques personnelles se sont multipliés, confrontant dans le même temps l'utilisateur aux problèmes liés à l'archivage et à l'accès à de grandes quantités de données. La démocratisation de l'accès à internet permet depuis quelques années l'échange de ces données via les sites spécialisés, menant à une augmentation inéluctable du volume des collections personnelles.

La possession et la conservation de ces banques de données ne présentent un intérêt que si nous sommes en mesure d'accéder efficacement à l'information qu'elles contiennent. C'est un enjeu majeur qui est à l'origine notamment du domaine de recherche dans lequel cette thèse s'inscrit.

1 L'indexation et la structuration de documents

1.1 Définitions générales

Un document est *« une unité représentant une contribution intellectuelle identifiée et publiée sur un média pour des raisons spécifiques. Un document exhibe, dans une certaine limite, une structure intentionnelle qui définit comment les éléments de son contenu sont organisés selon des axes (temps et espace) dans l'objectif d'être interprété par un lecteur comme témoignage de cet objectif original de publication. »* d'après la définition générale proposée dans [Auffret 99].

Dans le cas de documents analogiques, comme un livre ou une photographie, le lien entre l'objet physique et l'unité de document est assez direct. Il n'en est pas de même pour les documents numériques.

3. http ://www.ina.fr

La numérisation d'un document peut mener à sa fragmentation sur un disque dur. Les données binaires ne correspondent pas à un document comme il est expliqué dans [Troncy 04].
*« **Le document numérique** doit correspondre à ce qui est perçu par l'utilisateur grâce à un périphérique (écran, enceinte, imprimante). Les fragments d'un fichier stockés sur plusieurs pages d'un disque dur d'ordinateur ne sont pas des documents mais l'impression ou l'affichage du fichier dans une interface, une fois les opérations de réconstitution des différents fragments effectuées, en est un. »*

Les **documents audiovisuels** se distinguent des autres types de documents par leur nature temporelle. Techniquement, ces documents correspondent à un flux audiovisuel, c'est-à-dire une succession temporelle d'images superposées à un signal audio. La définition proposée par [Prié 99] précise qu'un document audiovisuel est avant tout une composition d'éléments réalisée par un auteur lui même contraint par un ensemble de règles de production audiovisuelle. Il précise que *« la diversité des types de documents audiovisuels, liée à leur production dans des objectifs, pour des publics et sous des formes différents en font un médium difficile à appréhender en soi, globalement, en tant que document »*.

Une collection de documents audiovisuels est un regroupement cohérent de documents vis-à-vis d'un critère physique, ou d'un critère sémantique. Pour établir une collection, il est nécessaire de posséder une description du contenu des documents. L'Ina emploie des documentalistes spécialisés dans la rédaction de notices documentaires décrivant ces contenus.

Comme les chaînes de télévision et les stations de radio diffusent des flux audiovisuels continus, l'annotation des documents audiovisuels a tout d'abord nécessité la définition d'une unité de base.

Cette unité est le programme audiovisuel. La définition proposée par [Manson 10] est assez précise. *« **Un programme** est un ensemble d'éléments consécutifs qui ne sont pas des inter-programmes et qui sont liés par une même charte audiovisuelle. Il est principalement à valeur culturelle, informative ou divertissante. Il peut être constitué de plusieurs parties séparées par des coupures publicitaires contenant des inter-programmes. Cela peut être un film, un épisode d'une série, un jeu, un journal, la météo, un clip, un magazine, un documentaire, etc. Chaque programme possède un titre qui le qualifie. Un flux télévisuel est composé de programmes entiers, de parties de programmes et d'inter-programmes assemblés consécutivement lors de la production du flux. »*

L'élaboration de notices documentaires précises et non ambiguës, pour chaque document archivé, est indispensable pour permettre une consultation efficace des archives. Cette documentation sur le contenu des documents relève du procédé d'indexation et de structuration du contenu que nous décrivons dans la suite.

1.2 Indexation de documents audiovisuels

L'indexation d'un document consiste à dégager les caractéristiques les plus représentatives de son contenu et à les représenter par un ensemble d'élément-clés (mots, images...).

Par exemple, Voxalead [4] est un moteur d'indexation permettant de rechercher du texte dans un contenu multimédia (audio ou audiovisuel). Les mots les plus utilisés sur les sites d'informations d'actualités sont indexés : une liste de mot-clés est proposée et des méta-données telles que la fréquence, ou la catégorie (people, organization, location) sont ajoutées par le biais de la mise page (cf. figure 1).

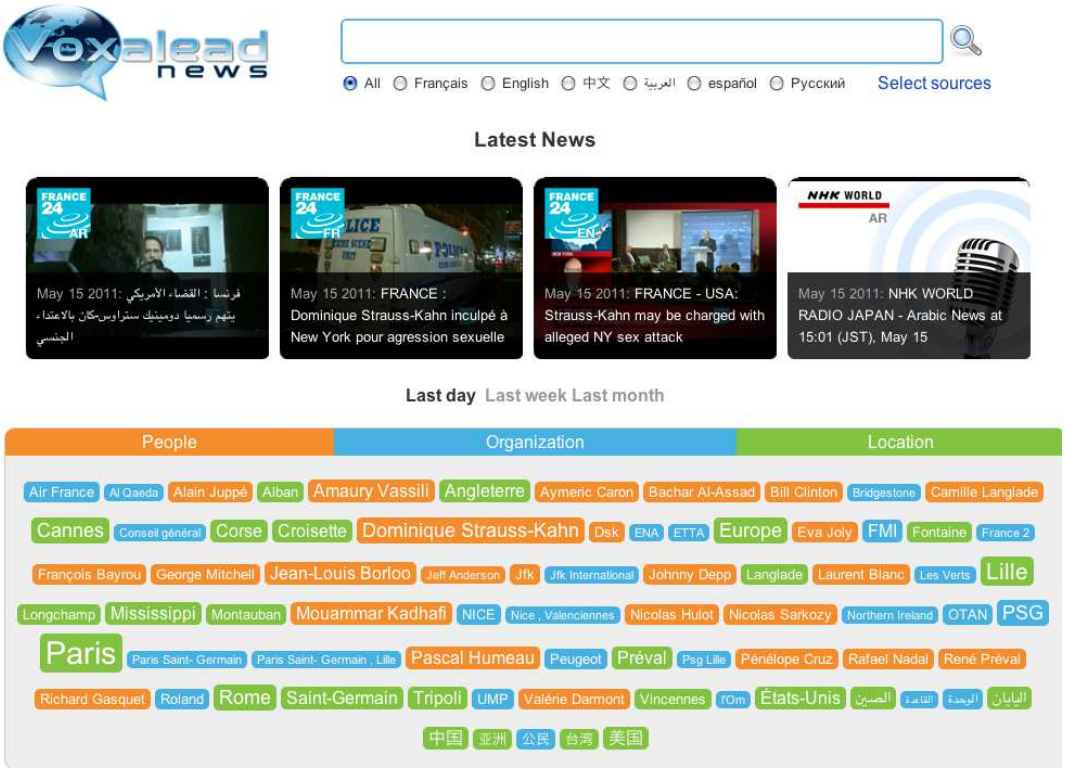


FIGURE 1 – Voxalead : un moteur d'indexation de contenus multimédias.

L'indexation d'un document nécessite une phase d'analyse ou d'interprétation du contenu du document. Le vocabulaire utilisé par les organismes officiels est codifié, ce qui permet au processus d'indexation de normaliser la codification du contenu des documents, et ainsi de faciliter la recherche du document par un utilisateur.

L'analyse du document est l'étape la plus sensible et la plus importante, sans laquelle il est impossible d'indexer correctement un document. Ce travail ne peut se résumer à une routine technique et fait appel à toutes les capacités intellectuelles des annotateurs. Il nécessite de parcourir le document, de ne pas se limiter au titre ou aux premières lignes de l'ouvrage, mais d'accéder à la table des matières, éventuellement lire la préface ou l'introduction. Si cela ne suffit pas, il faut parfois aller plus loin dans l'étude du document, il s'agit donc d'un travail difficile et très coûteux en temps [Baraggioli 08]. Une autre manière d'obtenir des informations sur le contenu d'un document est d'étudier sa structure.

4. http ://voxaleadnews.labs.exalead.com/

1.3 Structuration de documents audiovisuels

La structure d'un document peut se définir comme l'organisation des différents éléments de son contenu lui donnant sa cohérence et sa forme. La structure d'un document reflète l'intention première de son auteur et constitue donc une étape importante vers sa compréhension.

Par exemple, un journal télévisé se compose généralement de plusieurs grandes parties correspondant à de grandes thématiques comme la politique étrangère, la page des sports... Chaque grande thématique s'articule autour de quelques sujets, eux même structurés autour d'une introduction faite par le présentateur suivie d'une séquence de reportage développant le sujet. Un reportage peut à son tour contenir d'autres événements, comme une interview (cf. figure 2).

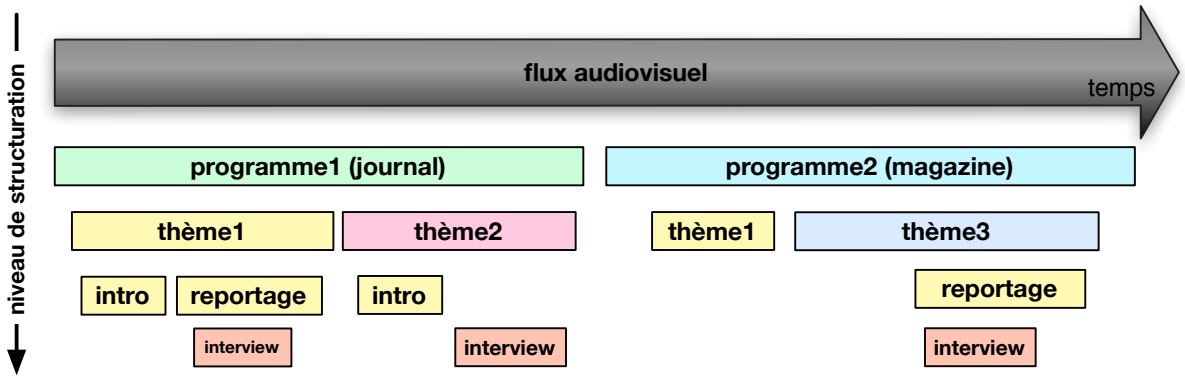


FIGURE 2 – Exemple de structuration d'un document audiovisuel (un journal suivi d'un magazine).

La structure d'un document audiovisuel peut donc être vue comme une table des matières du document. Elle permet d'avoir un aperçu rapide de son contenu à travers les événements extraits pour les différents niveaux de description du document. Les éléments de la table des matières peuvent concerner la forme (interview) ou le fond (thème).

Produire une annotation manuelle de qualité prend beaucoup de temps, or le volume de documents produits chaque année ne cesse d'augmenter. Il est indispensable de proposer des méthodes automatiques permettant d'extraire des caractéristiques sur de grandes quantités de documents.

1.4 Extraction automatique du contenu dans les documents audiovisuels

L'indexation et la structuration automatique des documents audiovisuels relèvent de l'extraction automatique d'éléments du contenu et de leur interprétation. Celles-ci reposent généralement sur l'extraction de paramètres numériques, calculés directement à partir des signaux audio et/ou vidéo du document numérique audiovisuel considéré. Ces paramètres visuels et acoustiques sont souvent très proches du signal, et une étape supplémentaire de description est nécessaire afin d'en dégager une sémantique exploitable dans une tâche d'indexation. En particulier, la norme MPEG-7 [Martinez 02] a pour ambition de proposer une normalisation des outils de description du contenu pour l'indexation et la recherche de documents numériques audiovisuels. Cette

normalisation des étapes de l'indexation de documents audiovisuels concerne à la fois les descripteurs, leurs méthodes d'extraction, ainsi que le langage de description. Descripteurs et langage de description sont souvent liés à la nature de l'objet d'interêt.

Un objet principal d'intérêt est l'activité humaine et nombre de travaux s'intéressent aux événements propres à décrire et à caractériser les **intervenants**. Le seul signal audio peut contenir des sons-clés caractéristiques d'intervenants humains comme des rires, des applaudissements. La détection de parole dans le flux audio plus particulièrement est essentielle :

- l'analyse de la parole permet une segmentation et un regroupement en locuteurs, suivie d'une éventuelle identification des locuteurs,
- la transcription de la parole permet d'extraire des informations sur les thématiques traitées,
- les descripteurs prosodiques, à travers l'intonation et le débit de parole peuvent être révélateurs d'intentions, d'opinions des intervenants.

La détection de la voix chantée conduit à la caractérisation audio du chanteur.

L'intervenant est également accessible à travers l'analyse du flux vidéo et des images grâce à des méthodes de détection des visages, de reconnaissance et de suivi des intervenants. Sur le même média, les textes incrustés peuvent contenir des informations sur l'identité ou la fonction des intervenants. Notons que l'analyse textuelle des sous-titres peuvent également permettre d'obtenir des informations sur les intervenants.

2 Problématique de recherche

À l'exemple du projet EPAC (qui sera présenté dans la section suivante), les enjeux, liés à l'indexation automatique des contenus audiovisuels, rassemblent vers un même objectif, des chercheurs issus de domaines différents, tels que le Traitement Automatique du Langage Naturel (TALN), le traitement numérique du signal audio, le Traitement et la Reconnaissance Automatique de la Parole (TAP et RAP). Le projet EPAC a plus particulièrement rassemblé les chercheurs de ces domaines dans le but de faire progresser les connaissances autour des difficultés actuelles rencontrées en traitement et en reconnaissance de la parole conversationnelle.

Les systèmes actuels de transcription automatique rencontrent des difficultés causées principalement par un ensemble de phénomènes acoustiques et langagiers apparaissant dans la parole produite dans un contexte conversationnel. Quand ils sont appliqués à des documents très structurés, contenant principalement de la parole lue ou préparée, comme dans des journaux d'informations, les systèmes de reconnaissance automatique de la parole atteignent des taux d'erreur-mot (WER) dans les meilleurs cas de l'ordre de 10%. Sur de la parole conversationnelle, le WER est beaucoup plus élevé : au-delà de 30%.

Les informations portées par la parole sont très utiles au processus d'indexation des documents audiovisuels. En effet, une partie des index est directement extraite des transcriptions automatiques analysées par des méthodes du domaine du TALN. Ces méthodes permettent, par exemple de détecter des mots-clés caractéristiques, des champs lexicaux exprimant une thématique ou des entités nommées (noms propres, lieux, dates...). Les documents sur lesquels le WER est faible peuvent être indexés directement à partir d'une analyse des résultats de transcription. Par contre, le même procédé d'indexation est à proscrire pour les documents contenant de la

parole conversationnelle : le WER peut être très élevé et mener à des index erronés. Or, dans une base de données immense (comme celle de l'Ina), un document mal indexé est un document perdu. De plus un document mal indexé pollue la base de documents en augmentant l'importance du bruit documentaire. Les conséquences sont telles que certains pourraient en conclure qu'il est préférable de ne pas indexer de cette manière ces documents pour le moment.

Toutefois, il n'existe, à notre connaissance, que peu de moyen automatique et efficace de détecter *a priori* si un document contient ou non de la parole conversationnelle. À la rigueur, les grilles des programmes peuvent fournir des indications sur les chances de rencontrer de la parole conversationnelle, notamment à travers les méta-données indiquant le genre du programme. Mais finalement, aucune information ne permet de localiser précisément les zones de parole conversationnelle à l'intérieur du flux audio.

Les travaux de recherche présentés dans ce manuscrit se rassemblent autour de cette problématique. Notre contribution réalise une structuration automatique des documents audiovisuels, et vise à faire émerger des zones particulières des documents correspondant à des conversations. Il est important que ce résultat soit indépendant du contenu du message transmis, de manière à ne pas intégrer dans cette détection les erreurs inhérentes à la reconnaissance automatique de la parole conversationnelle. Nos méthodes ont été développées en exploitant uniquement le flux audio des documents audiovisuels. Ce choix nous paraît cohérent car les conversations sont, dans les documents audiovisuels, des événements principalement sonores. Cette méthode de structuration peut ainsi être appliquée indifféremment sur le flux ou une partie du flux audio de documents télévisuels et radiophoniques quelle que soit la langue utilisée dans le document.

Notre méthode de structuration permet de localiser les bornes temporelles de début et de fin des conversations. Cette connaissance *a priori* permet d'adapter l'indexation automatique donnant la possibilité :

- de transcrire uniquement les zones non conversationnelles en vue d'indexer une partie ou la totalité du document,
- d'appliquer aux zones de conversations, une méthode de reconnaissance automatique adaptée à la parole conversationnelle,
- de proposer la réalisation d'une transcription manuelle de la séquence conversationnelle,
- éventuellement, de ne pas traiter le document s'il contient une trop grande proportion de parole conversationnelle,
- de collecter un ensemble de données spécifiques à l'étude des phénomènes linguistiques et non linguistiques mis en jeu par les locuteurs dans un contexte conversationnel...

Notre intuition est également que les zones de conversation, sont intrinsèquement porteuses d'une grande quantité d'informations sur la nature du document. Nous proposons dans ce sens de caractériser les conversations afin de faire ressortir leurs particularités. Nous nous sommes imposé les deux contraintes suivantes :

- la caractérisation des zones de conversation doit être indépendante de la transcription. Ceci afin de s'affranchir des erreurs inhérentes aux méthodes automatiques sur ce type de contenu.

– les zones de conversation pouvant apparaître sur une grande variété de documents, leur caractérisation doit être indépendante du type de programme et de la structure du programme.

Nous avons choisi de développer une méthode de caractérisation des zones de conversation en fonction des rôles des intervenants. Les rôles - présentateur, journaliste et invité - sont suffisamment génériques pour correspondre à un grand nombre de programmes différents. Dans le même temps, cette information peut être utilisée pour caractériser les conversations dans la mesure où une discussion entre un présentateur et un invité n'a souvent pas le même objectif qu'une discussion entre un présentateur et un journaliste.

Des méthodes particulières d'extraction d'informations pourront alors être appliquées à chacune des catégories de conversation afin d'extraire de manière ciblée l'identité de la personne interviewée ou la thématique de la chronique.

3 Contexte scientifique et contractuel de l'étude

Les travaux effectués dans cette thèse s'insèrent au cœur des sujets de recherche de l'équipe SAMoVA (Structuration, Analyse et Modélisation de documents Vidéo et Audio) de l'IRIT (Institut de Recherche en Informatique de Toulouse). Les compétences en segmentation du signal et en traitement de la parole ont été mises au service des problèmes d'indexation et de structuration du flux audiovisuel, avec pour exemples :

– la segmentation du signal audio en ses composantes primaires (parole, musique, bruit) [Pinquier 04]. Ce travail a été complété par une caractérisation de l'environnement musical [Lachambre 09].
– la recherche des locuteurs dans le flux audio avec les études en segmentation et regroupement en locuteurs [El Khoury 10] et en vérification du locuteur [Louradour 07],
– la caractérisation des relations temporelles entre locuteurs [Ibrahim 07],
– la détection de mots-clés [Le Blouch 09].

D'une part, la contribution scientifique présentée dans ce manuscrit tire bénéfice des travaux antérieurs, axés sur la détection et la caractérisation des intervenants dans les contenus audio et vidéo. D'autre part, la problématique de la parole conversationnelle et la structuration de documents sont des objets d'études partagés par de nombreux laboratoires en parole, d'où la proposition du projet EPAC « Exploration de masses de documents audio pour l'extraction et le traitement de la PArole Conversationnelle » auquel s'est associée l'équipe SAMoVA.

Le projet EPAC [Estève 10] est une des réponses à l'appel d'offres Masse de Données lancé par l'Agence Nationale de la Recherche, et financé de janvier 2007 à août 2010. Il a réuni quatre partenaires, qui sont le LI (Laboratoire d'Informatique de Tours), le LIA (Laboratoire d'Informatique d'Avignon), le LIUM (Laboratoire d'Informatique de l'Université du Maine), porteur du projet, et l'IRIT dont une des contributions est rapportée dans ce manuscrit. *Le projet EPAC a eu pour but de proposer des méthodes d'extraction d'information et de structuration de documents, spécifiques aux données audio, prenant en compte l'ensemble des canaux d'information :*

segmentation du signal (parole/musique/jingle/...), identification et suivi du locuteur, transcription de parole, détection et suivi de thème, détection d'entités nommées, analyse du discours, interactions conversationnelles, etc.

À l'issue de ce projet, a été fourni un corpus audio de 100 heures, totalement transcrites et annotées : 10 heures ont été traitées manuellement et le reste de manière semi-automatique. Un accent particulier a été porté sur les zones de parole conversationnelle et les zones de parole spontanée. Ce corpus rassemble des émissions de France Info, France Culture et RFI, diffusées entre 2003 et 2004. Finalement, les sorties automatiques produites par les différents outils des partenaires du projet EPAC pour l'ensemble des 1500 heures d'audio brut de ESTER 1 [Galliano 05] sont venues s'ajouter au corpus final. La contribution de l'équipe SAMoVA s'est portée sur deux aspects :

- la segmentation de base du signal audio et la localisation des locuteurs (segmentation et regroupement automatique),
- la structuration des documents en type de zones d'interaction (type de locuteur et nature de l'interaction).

Cette thèse étant financée par le projet, une grande partie des travaux réalisés s'insère dans ce cadre, néanmoins la problématique a été élargie.

4 Plan du manuscrit

Le chapitre 1 de ce manuscrit est nommé *Étude préliminaire sur la détection et la caractérisation des zones d'interaction orale entre intervenants*. Ce chapitre présente les premiers pas de nos travaux autour de la détection des zones sonores susceptibles de contenir de la parole conversationnelle. Ce travail s'ouvre sur une première catégorisation des locuteurs qui sert de base à notre contribution au domaine de la reconnaissance automatique des rôles.

Le chapitre 2 est intitulé *Des paramètres pertinents pour la reconnaissance automatique des rôles*. Il s'agit de la première partie de notre contribution à la reconnaissance des rôles des locuteurs. Nous y présentons un ensemble de paramètres simples à extraire du signal audio et des tours de parole des locuteurs. La pertinence des descripteurs proposés est évaluée avant de les intégrer dans un système complet présenté dans le chapitre suivant.

Le chapitre 3 – *Système de reconnaissance automatique des rôles* – est dédié à l'étude expérimentale d'un système complet fondé sur l'architecture classique d'un système de reconnaissance des formes. Un grand nombre d'expériences sont réalisées dans le but d'étudier l'influence de chacune des étapes du système et d'établir la meilleure configuration.

Le chapitre 4 se nomme *Structuration de documents audiovisuels et rôles des intervenants*. Nous y présentons une méthode de structuration fondée sur la connaissance des rôles des intervenants et des zones d'interaction. Cette étape de notre travail intègre les systèmes développés et présentés dans les chapitres précédents. Nous refermons ensuite ce manuscrit par un chapitre de conclusions et de perspectives.

Chapitre 1

Étude préliminaire sur la détection et la caractérisation des zones d'interaction orale entre intervenants

Dans les documents audiovisuels, les interventions des locuteurs peuvent correspondre à des *dialogues*, ou à des *monologues*. Comme il est expliqué dans [Kerbrat-Orecchioni 95], les termes *dialogue* et *monologue* peuvent avoir plusieurs définitions.

- Le dialogue :
  - *au sens strict*, implique des interventions verbales alternatives entre deux locuteurs physiquement distincts ; on dira de ce discours échangé qu'il est de nature dialogale.

  - *au sens étendu*, est un discours adressé, mais qui n'attend pas de réponse, du fait du dispositif énonciatif dans lequel il s'inscrit. Ces discours unilatéraux peuvent être dialogiques dans la mesure où ils incorporent plusieurs voix, imputables à autant d'énonciateurs distincts. Cette forme de dialogue correspond aux locuteurs qui ne s'écoutent pas mutuellement, cette situation pouvant conduire parfois à de la parole recouvrante.

- le monologue :
  - *au sens strict*, est un discours non adressé, si ce n'est à soi-même.
  - *au sens étendu*, est un discours adressé à une audience, mais qui ne permet pas l'alternance.

Dans le cadre de nos travaux, un *dialogue* répondra aux 2 définitions, stricte et étendue, qui se rejoignent sur la notion d'une alternance des interventions de deux locuteurs distincts. Dans des émissions de radio ou de télévision, ces événements peuvent correspondre à toutes formes de discussion telle une interview ou un débat.

Nous retiendrons la seconde définition du *monologue*, comme étant celle que nous rencontrerons dans notre étude. La lecture des titres par un présentateur de bulletin d'information est un monologue d'après cette définition ; le message étant adressé aux spectateurs.

Un dialogue peut correspondre à plusieurs types d'interactions verbales (*verbal* est utilisé ici dans le sens de *langagier*, *linguistique*). Il peut s'agir d'un entretien, d'un débat, d'une négociation, etc. Ces interactions verbales se distinguent d'une conversation, qui en est la forme la plus commune et la moins formatée. Une conversation a un caractère familier, improvisé et n'ayant pas de but en soit, si ce n'est le plaisir de parler [André-Larochebouvy 84].

Dans ce chapitre nous présentons une étude centrée sur la détection automatique des interactions verbales entre deux intervenants. Dans notre approche, nous n'utilisons pas de connaissances sur le contenu linguistique des interventions des locuteurs. En effet, peu importe le message, l'information que nous utilisons indique seulement si un locuteur parle ou ne parle pas à un instant donné. Pour mettre en relief cette caractéristique nous utiliserons le terme **d'interaction orale**, plutôt que verbale. Nous considérons indifféremment les définitions strictes et étendues d'un dialogue. Nous n'avons pas non plus d'*a priori* sur la nature des interactions (conversation, débat...).

Une **zone d'interaction orale** est une zone temporelle d'un document durant laquelle deux locuteurs font alternativement des interventions détectées comme correspondant à de la parole. Nous nous affranchissons d'un *a priori* sur le contenu linguistique, et nous nous assurons que la structure de la zone respecte les définitions énoncées.

Ce travail est motivé d'une part par la nécessité de trouver des éléments structurants, caractéristiques du contenu des documents et d'autre part par la recherche de zones contenant potentiellement de la parole spontanée. Nous présentons ensuite quelques éléments de l'état de l'art qui nous permettent de situer notre contribution (cf. section 1.2). Dans la section 1.3, nous décrivons la méthode nous permettant de détecter des zones dites d'interaction orale entre deux locuteurs ainsi que l'étude menée pour caractériser plus précisément la nature de leur contenu. Ceci sera le point de départ des travaux menés sur la reconnaissance des rôles des locuteurs traitée dans le chapitre suivant.

1.1 Motivations

1.1.1 La structuration par le contenu de documents audiovisuels

La grille des programmes de radio et de télévision (au sens de l'EPG [5]) contient la suite des programmes diffusés. Pour chaque diffusion, c'est-à-dire chaque instance d'un programme, que nous appellerons « émission », il est indiqué son heure de diffusion, son éventuelle durée, le type (journal, divertissement, documentaire, film...) et parfois l'identité des intervenants ou le thème abordé. La connaissance *a priori* du type d'émission peut fournir une indication sur la possibilité d'y trouver des séquences conversationnelles. D'une part, cette information est incertaine. D'autre part, la grille des programmes n'a pas le niveau de détail nécessaire afin de localiser les zones conversationnelles dans l'émission.

Pour détecter automatiquement dans ces documents les passages susceptibles de contenir des interactions orales entre plusieurs intervenants, il est nécessaire de proposer une méthode d'analyse par le contenu audio et/ou vidéo. Elle permettra de localiser temporellement les zones

5. Electronic Program Guide

de dialogue et de monologue d'un document audiovisuel, comme nous l'avons représenté sur la figure 1.1 pour l'exemple d'un bulletin d'information. Sur cette figure, la première ligne correspond aux différentes séquences réelles du journal (titres, sujets, interviews, résumé des titres). Sur la seconde ligne, nous avons représenté le résultat que fournirait une détection idéale des zones d'interaction. Cette représentation fait émerger des informations sur la structure interne du document. La détection des deux zones d'interaction conduisent à la localisation des interviews dans ce type d'émission.

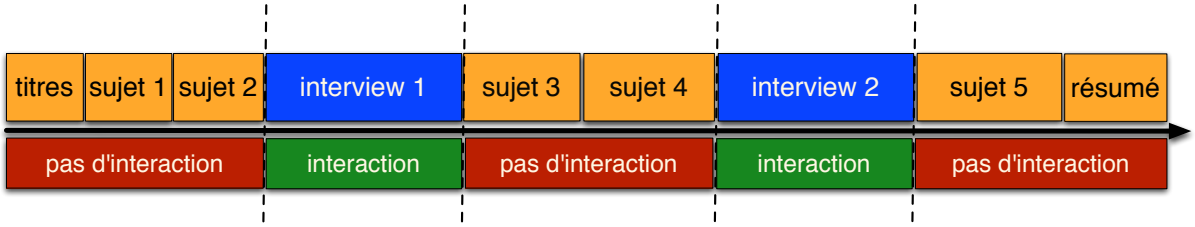


FIGURE 1.1 – Séquences d'un journal (frise du haut) et résultat possible d'une détection automatique des zones d'interaction (frise du bas).

1.1.2 La détection des zones contenant de la parole conversationnelle

Outre son intérêt d'un point de vue « structuration » d'un document, la détection des zones d'interaction peut aider à la localisation des zones de parole spontanée, en opposition à celles de parole préparée. Dans notre étude, la parole spontanée [Luzzati 04] doit être entendue dans le sens « d'un énoncé perçu et conçu au fil de son énonciation ».

Les récentes campagnes d'évaluation ESTER[6] [Galliano 05] et ESTER2 [Galliano 09] rapportent les bonnes performances des méthodes actuelles de reconnaissance automatique de la parole sur des documents contenant de la parole préparée. Le meilleur système présente un taux d'erreur mot (Word Error Rate) de l'ordre de 11%. Ces scores obtenus sur des corpus francophones se rapprochent des performances atteintes en langue anglaise malgré les difficultés inhérentes à la langue française décrites dans [Gauvain 05a]. La parole utilisée par les locuteurs durant une conversation n'est pas systématiquement spontanée, mais la nature improvisée et le cadre peu formaté de ces interactions est propice à la production de ce type de parole. Comme le rapporte [Gauvain 05b] : *Les difficultés posées par la modélisation linguistique de la parole conversationnelle résident d'une part dans le caractère spontané de la parole qui conduit à une syntaxe relâchée avec de nombreuses hésitations et reprises, et d'autre part dans la faible quantité de données d'apprentissage disponible.* Ces observations résonnent avec celles faites lors de l'étude menée par [Furui 05] sur un grand corpus de parole spontanée japonaise : *[...] acoustic and linguistic variation of spontaneous speech is so large that we need a very large corpus in order to encompass the variations.* La spontanéité du langage au sein d'une zone d'interaction orale s'accroît naturellement avec le degré d'interaction ; or c'est un facteur d'erreurs en transcription automatique qu'il est intéressant de pouvoir détecter en amont de la reconnaissance de la parole afin, par exemple, de spécialiser les données d'apprentissage.

6. Évaluation des Systèmes de Transcription Enrichie d'émissions Radiophoniques

Dans [Dufour 09], les auteurs évaluent également le lien entre le degré de spontanéité du langage et les performances de reconnaissance. Ils concluent sur l'intérêt d'une connaissance *a priori* du degré de spontanéité en perspective d'une amélioration des systèmes de transcription automatique de la parole conversationnelle. Dans cette optique, nous souhaitons contribuer en proposant une approche non linguistique pour classer des séquences conversationnelles en fonction de leur degré d'interactivité, et ainsi aider à la prédiction de la spontanéité. Les travaux de [Burger 02] vont dans ce sens. Il y est tenté d'établir un lien entre le type de réunion (professionnel, bavardage...) et le type de parole utilisée (spontanée ou préparée). Les résultats montrent que bien qu'il existe des phénomènes linguistiques liés au type de réunion, d'autres informations concernant les intervenants (âge, liens sociaux entre les participants) pourraient être nécessaires à la catégorisation du type de parole.

1.2 État de l'art

Les interactions sociales et verbales entre individus sont largement étudiées depuis les années 50 dans le domaine de la sociologie, de l'ethnologie et de la psychologie [Bale 50]. L' « analyse conversationnelle » est axée sur l'étude des interactions entre individus dans des situations courantes. En 1974, les travaux de [Sacks 74] établissent par l'observation un ensemble de règles régissant l'organisation temporelle des tours de parole entre locuteurs impliqués dans une conversation. Ces résultats sont fondamentaux et cités encore aujourd'hui à une large échelle dans des domaines variés. La détection des conversations est étudiée plus récemment en informatique sous différents angles que nous allons couvrir dans cet état de l'art.

1.2.1 Reconnaissance automatique des actes de dialogue

Un acte de dialogue est lié à la sémantique d'une phrase ou d'une pseudo-phrase prononcée par un locuteur durant une conversation. Les travaux de [Stolcke 00] et [Shriberg 98] proposent initialement 42 actes de dialogue différents mais cet ensemble est souvent réduit aux actes de dialogue suivants : énoncé, question, reformulation, phrase incomplète, agrément, appréciation, autres. Les actes de dialogue sont modélisés, puis utilisés pour faire apparaître la structure de la conversation. La structure est révélatrice de l'état d'avancement de la compréhension du déroulement du dialogue et ainsi sa connaissance peut améliorer les performances des systèmes de reconnaissance automatique de la parole conversationnelle. Des informations prosodiques peuvent être combinées aux informations linguistiques pour rendre plus robustes les modèles.

Les travaux de [Kral 05] sont orientés vers la mise en œuvre d'un système de réservation automatique et de l'animation d'une tête parlante. Ces applications nécessitent de reconnaître des actes de dialogue dans un énoncé (ordres, questions fermées et ouvertes). À partir d'informations lexicales et prosodiques, et d'une approche fondée sur les réseaux de neurones, les auteurs parviennent à classer correctement 92% des actes de dialogue de l'ensemble de test.

1.2.2 Détection des interactions orales dans les groupes

Dans les travaux en détection des conversations dans les groupes d'individus (cocktail party, réunions), un soin tout particulier est apporté au mode de capture des signaux. Les intervenants

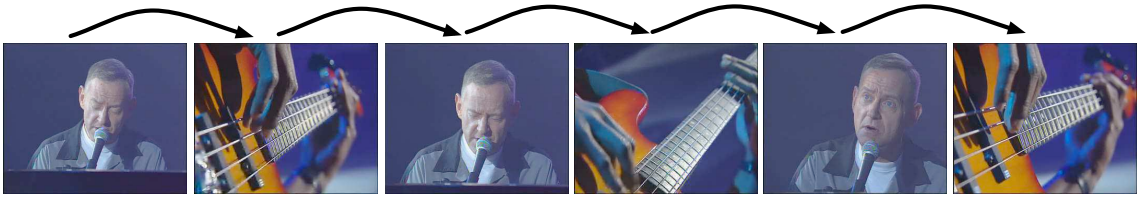
sont alors équipés de microphones portables individuels de telle sorte qu'il existe K signaux audio pour un groupe de K individus. Dans un premier type de système [Corman 94], il est supposé que lorsque deux personnes vont se parler face-à-face, leurs microphones respectifs vont enregistrer des signaux audio semblables contenant les contributions des deux locuteurs. Les paires de locuteurs qui interagissent sont détectés à l'aide du calcul de la corrélation entre les signaux. Dans un second type d'approche diamétralement opposée [Basu 02, Wyatt 07], l'enregistrement de chaque locuteur est analysé par un détecteur d'activité vocale afin d'isoler les instants de parole du porteur du microphone. Le calcul de l'information mutuelle des signaux permet de détecter les couples de locuteurs qui conversent. Par cette dernière méthode il est possible de détecter plusieurs conversations ayant lieu simultanément [Brdiczka 05].

L'acquisition des signaux audio est ici très contrainte. Couramment un seul signal audio mono ou stéréo contenant les contributions de tous les intervenants est disponible comme c'est le cas lors d'enregistrements télévisuels et cinématographiques.

1.2.3 Détection des scènes de dialogue dans les documents vidéo

Dès lors que les documents sont audiovisuels, le terme « acte de dialogue » fait place à celui de « scène de dialogue ». Le contenu de tels documents est classiquement découpé en scènes, comme les scènes d'action, de transition, de description, de dialogues... Cette catégorisation se fait principalement à partir d'informations visuelles.

(a) Une séquence d'images correspondant à un dialogue



(b) Une séquence d'image ne correspondant pas à un dialogue

FIGURE 1.2 – Deux séquences de plans illustrant la définition visuelle d'une scène de dialogue.

Plusieurs méthodes [Wolf 97, Eickeler 99, Ferman 99] utilisent une approche supervisée pour la reconnaissances des types de scène. D'autres contributions [Saraceno 98, Sudaram 02, Zhai 04] sont quant à elles fondées sur un ensemble de règles. Une scène de dialogue dans ces travaux se définit comme une séquence d'alternances de plans vidéo redondants. Cette définition atteint rapidement ses limites puisque, comme nous le représentons dans la figure 1.2, dans un même document une séquence d'alternances peut correspondre à une conversation 1.2(a) ou à une séquence d'une toute autre nature 1.2(b). Dans ces travaux, l'information audio n'est pas utilisée en dépit du fait que les conversations soient des événements oraux et que cette information reste

certainement la plus pertinente. Dans la suite nous nous focalisons sur des approches favorisant l'exploitation des informations extraites de l'audio.

1.2.4 Détection des interactions par une prise en compte de l'audio

Dans les travaux de [Aydin Alatan 01], bien que les données traitées soient issues d'un corpus de séries télévisées (sitcoms et téléfilms), la présence de parole est considérée comme fondamentale à la définition d'une scène de dialogue. Ce point distingue cette contribution des approches uniquement fondées sur des critères visuels. Pour l'auteur, une scène de dialogue doit regrouper trois événements qui sont par ordre d'importance : la présence de parole, la présence d'un visage et un continuum de lieu. Pour chaque plan trois événements sonores (parole, musique et silence) sont détectés par analyse fréquentielle et par analyse de l'énergie du signal audio. La présence ou l'absence d'un visage dans le plan est identifiée suite à une détection de la couleur de la peau. Un changement de lieu est détecté par une modification significative des histogrammes de couleurs du plan courant par rapport au plan précédent.

Le système de localisation des scènes de dialogues utilise des modèles de Markov cachés (MMC). Une scène de dialogue correspond dans ce cas à un état du MMC. Les performances obtenues en fusionnant l'audio, le visage et le changement de lieu atteignent 85% des scènes bien reconnues. La fusion de l'audio et du visage permet de reconnaître 90% des scènes de dialogue. L'audio seul atteint 75% de scènes bien reconnues. Ce dernier résultat est présenté comme un très bon compromis (temps de calcul/performances) en comparaison du meilleur système. Notons que dans cette étude les données de test sont intégrées aux données d'apprentissage et que d'autre part, ces scores doivent être relativisés car dans les séries télévisés, une large proportion de la parole correspond à des scènes de dialogue. Néanmoins ces résultats montrent que l'audio peut être suffisant pour détecter les dialogues dans des documents audiovisuels.

1.2.5 Caractérisation des scènes conversationnelles

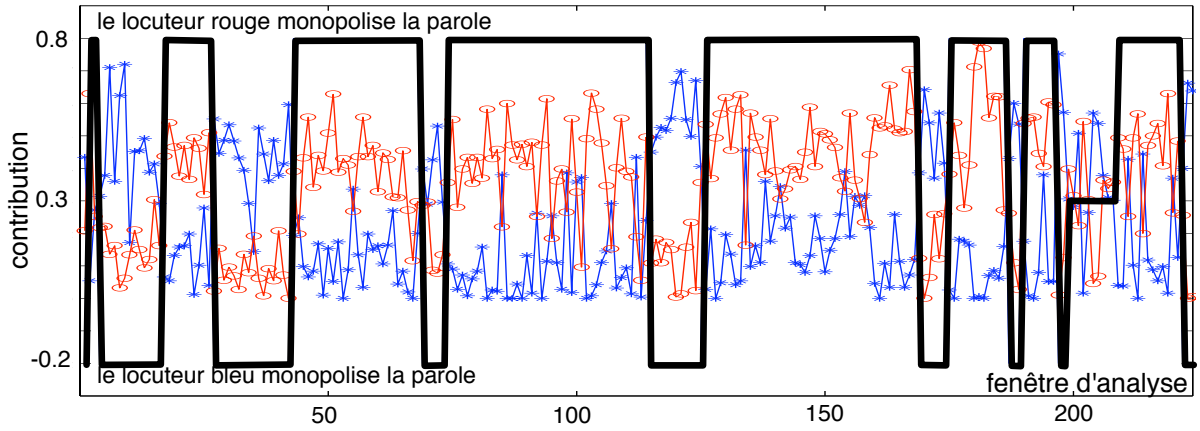
Pour être complète, la détection des conversations doit être accompagnée d'une description du contenu des séquences détectées. Nous présentons une contribution intéressante sur la caractérisation de séquences conversationnelles, celle de [Basu 02]. Elle repose sur l'observation qu'au même titre qu'en vidéo, une conversation peut être divisée en plusieurs scènes conversationnelles en fonction de l'évolution du contenu de la conversation.

Des conversations téléphoniques d'une demie-heure entre des locuteurs anglophones sont analysées. Une conversation téléphonique est composée de deux voies sonores temporellement concurrentes, correspondant à l'enregistrement des deux locuteurs. Dans ce cas, les intervenants ne sont pas en contact visuel et une conversation peut contenir des zones de parole superposée.

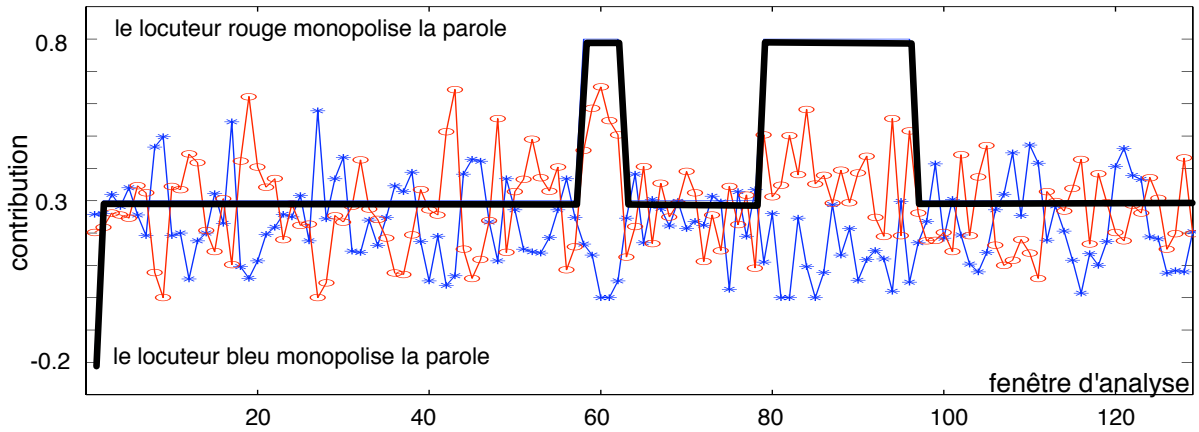
La première étape est la segmentation des conversations en scènes conversationnelles. La définition d'une scène conversationnelle repose sur l'observation que dans la majorité des conversations, l'un ou l'autre des locuteurs monopolise la parole. Le locuteur dominant peut alterner au cours de la conversation. Trois types de scènes conversationnelles sont proposées :

1. les scènes où le locuteur 1 monopolise la parole,
2. les scènes où le locuteur 2 monopolise la parole,
3. les scènes où il n'y a pas clairement de locuteur dominant.

La classification en scènes conversationnelles utilise un MMC à trois états. Les observations sont les proportions de temps de parole des deux locuteurs, calculées sur une fenêtre glissante de 8 secondes. Finalement 96% des scènes conversationnelles sont correctement affectées.



(a) Une conversation durant laquelle le locuteur dominant change souvent implique de nombreuses scènes conversationnelles.



(b) Une conversation durant laquelle il y a peu de zones présentant un locuteur dominant.

FIGURE 1.3 – Deux segmentations en scènes conversationnelles, *extrait de [Basu 02]*

Les courbes de la figure 1.3 sont extraites de [Basu 02], elles représentent les résultats de la segmentation en scènes conversationnelles pour deux conversations différentes. Sur ces figures, les courbes discontinues rouges et bleues représentent les contributions des temps de parole des locuteurs sur chaque fenêtre d'analyse. La courbe noire en créneaux indique le type de scène conversationnelle détectée. Cette courbe est en position haute pour indiquer que le locuteur rouge monopolise la parole. *A contrario*, cette courbe est en bas quand le locuteur bleu monopolise la parole. Enfin cette courbe est en position centrale pour indiquer qu'aucun des locuteurs ne monopolise la parole. La sous-figure 1.3(a) correspond une conversation durant laquelle le rôle de locuteur dominant alterne beaucoup. La sous-figure 1.3(b) présente la segmentation d'une conversation durant laquelle il n'y a pas de locuteur dominant hormis sur deux scènes.

Cette segmentation est utilisée pour catégoriser les conversations en fonction du niveau de dominance globale et de la longueur moyenne des scènes conversationnelles. Trois types de conversations sont ainsi mis en évidence. D'abord les conversations de type « bavardage », pour des valeurs faibles de la dominance et des scènes courtes. Les conversations de type « conférence » avec une grande dominance et de longues scènes, et les conversations de type « négociation » avec peu de dominance et des longues scènes conversationnelles.

Cette classification semble pertinente car elle repose sur des observations simples qui peuvent être appliquées de la même manière à des conversations d'une autre nature telles que des interviews ou des débats.

1.2.6 Synthèse de l'état de l'art

Les travaux de la littérature concernant l'étude des interactions et des conversations dans les documents audiovisuels rassemblent une large variété d'approches.

Des travaux se basent sur les transcriptions pour reconnaître les actes de dialogue caractéristiques d'une interaction orale. Cette approche nécessite, pour être performante, d'utiliser des transcriptions textuelles de bonne qualité dans un contexte conversationnel difficile pour les systèmes de reconnaissance automatique. D'autres approches utilisent uniquement des paramètres visuels. Ce choix limite l'utilisation de ces méthodes à des documents vidéo. De plus, les conversations étant par nature des événements liés à la parole, l'utilisation de paramètres visuels semble inadaptée ou du moins incomplète à la détection des conversations. Une troisième catégorie de travaux exploitent des informations extraites de l'audio et repose sur des hypothèses fortes sur l'acquisition des signaux de parole en nécessitant l'enregistrement individuel de l'activité vocale de chaque locuteur. Cette contrainte ne convient pas à la majorité de documents audiovisuels.

Nous souhaitons faire évoluer la définition d'une scène de dialogue en proposant de réaliser la détection d'alternances de tours de parole entre deux locuteurs différents. Nous ajoutons pour cela une information sur l'identité du locuteur. Nous voulons montrer qu'il est possible de faire émerger les séquences d'alternances de tours de parole entre deux intervenants uniquement à partir du signal sonore.

1.3 Notre approche pour la détection et la caractérisation des zones d'interaction

La contribution de [Ibrahim 07], réalisée dans notre équipe, a servi de base à notre travail en mettant en évidence l'existence de classes de relations temporelles caractéristiques d'interactions orales entre les tours de parole des locuteurs. Un résumé en est donné ci-après (section 1.3.1).

Ce travail antérieur à notre étude, ne concerne pas directement la localisation des conversations mais permet de vérifier l'existence dans un document d'une séquence :

$$Loc_1 - Loc_2 - Loc_1 - \dots$$

où Loc_x signifie le locuteur x. C'est à partir de ce résultat que nous fondons notre contribution sur la détection des zones d'interaction.

1.3.1 Relations temporelles entre locuteurs

[Ibrahim 07] propose une méthode générique de caractérisation des structures des documents audiovisuels fondée sur l'analyse des relations temporelles entre des caractéristiques (audio et vidéo) extraites d'un même document. Une caractéristique est représentée par un ensemble de segments temporels indiquant à quels instants (de début et de fin) elle se produit. L'auteur propose de calculer sur ces ensembles de segments plusieurs paramètres représentant les relations temporelles liant les caractéristiques. Comme nous l'avons représenté sur la figure 1.4, pour une paire de segments (s_{C1}, s_{C2}) correspondant respectivement à une caractéristique $C1$ et $C2$, il est possible de calculer trois paramètres (DE, DB et Lap).

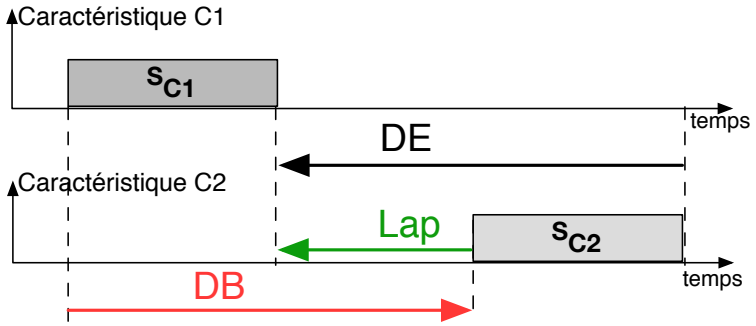


FIGURE 1.4 – Extraction des paramètres temporels entre deux segments, *extrait de [Ibrahim 07]*.

Le paramètre DE est la durée séparant la fin du segment S_{C2} et la fin du segment S_{C1}. Le paramètre DB est la durée entre le début du segment S_{C1} et le début du segment S_{C2}. Le dernier paramètre Lap est la durée séparant le début de S_{C2} de la fin de S_{C1}. Ces calculs sont étendus à l'ensemble des paires de segments distants d'une durée inférieure à un seuil α tel que $|Lap| \leq \alpha$. Tous les triplets de paramètres (DE, DB, Lap) d'un couple de caractéristiques $(C1, C2)$ sont rassemblés dans une matrice dite des relations temporelles. L'application, sur la matrice, d'une méthode de regroupement non supervisé (algorithme des K-means) permet de mettre à jour plusieurs classes de relations temporelles propres au couple de caractéristiques $C1$ et $C2$. Une expérience en particulier a été menée sur des caractéristiques correspondant aux tours de parole des locuteurs d'un document. Cette méthode a fait émerger du contenu audio les relations temporelles mises en jeu entre des locuteurs impliqués dans une conversation.

1.3.2 Détection enrichie des zones d'interaction orale

1.3.2.1 Présentation de l'approche

Notre méthode de détection des conversations est fondée sur la recherche de séquences d'alternances des tours de parole de deux locuteurs. Nous ne recherchons pas directement les conversations entières mais plutôt des zones d'interaction correspondant à des séquences d'alternances de locuteurs les plus longues possible.

Les documents audiovisuels, correspondant à des émissions de télévision et de radio, peuvent contenir des événements sonores variés comme des rires, des applaudissements, de la musique,

etc. Ceux-ci peuvent s'intercaler dans une séquence d'alternances et rendre difficile la détection de conversations entières. La recherche de zones d'interaction semble plus robuste dans ce cas.

La figure 1.5 présente les modules de notre système de détection et de caractérisation des Zones d'Interaction (Z.I.).

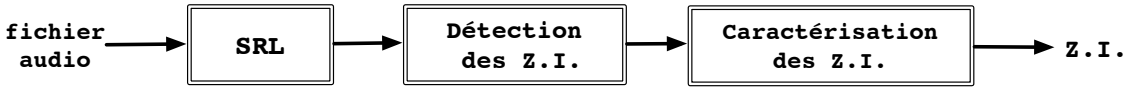


FIGURE 1.5 – Système de détection et de caractérisation des zones d'interaction.

L'approche proposée repose sur trois étapes :

- le premier traitement consiste à segmenter le signal audio en tours de parole ; un tour de parole correspond à un changement de locuteur. Cette segmentation est accompagnée d'une phase de regroupement en locuteurs afin de spécifier le nombre de locuteurs et l'affectation de chaque segment à un identifiant de locuteur (cf. section 1.3.2.2).
- le deuxième traitement a pour but de localiser les zones d'interaction (cf. section 1.3.3) indirectement à travers l'analyse des intervenants qui y sont impliqués. Nous calculons dans ce but des paramètres qui caractérisent globalement les interventions des locuteurs dans le document. Nous nous concentrons également sur la caractérisation de l'intervention du locuteur dans la zone d'interaction.
  Pour chaque zone d'interaction détectée, nous proposons une mesure nommée « le niveau d'interactivité », indiquant la longueur de la séquence d'alternance des tours de parole.
- le dernier traitement concerne la caractérisation des zones d'interaction grâce à des informations liées aux locuteurs (cf. section 1.3.4).

1.3.2.2 Segmentation et Regroupement en Locuteurs (SRL)

La SRL est un traitement de la parole qui consiste en une détection des tours de parole et leur assignation à une étiquette correspondant à un locuteur. Généralement ces systèmes procèdent sans connaître *a priori* le nombre de locuteurs présents dans le document. Dans une première étape, le contenu audio est segmenté en zones homogènes d'un point de vue acoustique. La seconde étape correspond au regroupement des segments. L'objectif est de rassembler les segments similaires dans un cluster correspondant à un locuteur.

Dans notre étude, nous utilisons l'outil de SRL développé dans notre équipe de recherche par El Khoury [El Khoury 10]. Les trois étapes de la méthode (détection de la parole, segmentation en locuteurs et regroupement) sont intégrées dans un processus itératif afin d'améliorer les performances sur les zones où plusieurs types de sources sont présentes. Les paramètres extraits du signal audio sont des coefficients MFCC et la modulation de l'énergie à 4 Hertz. L'étape de segmentation en locuteurs est basée sur le calcul du Maximum de Vraisemblance Généralisé (GLR). Le Critère d'Information Bayésienne (BIC) est ensuite exploité pour détecter les instants de changement de locuteurs. L'utilisation conjointe de la segmentation bidirectionnelle

(forward-backward) GLR/BIC augmente la robustesse de la méthode en détectant des transitions entre segments de locuteurs qui n'aurait pas été trouvées en une seule passe. Une approche ascendante de regroupement hiérarchique permet de rassembler les segments appartenant à un même locuteur. Même si la méthode nous assure que les segments les plus longs et les plus homogènes permettent un meilleur regroupement, dans le cas de la parole conversationnelle les segments tendent à être plus petits et à contenir de la parole superposée. Pour améliorer la phase de regroupement, El Khoury propose d'appliquer un regroupement local appliqué tous les 20 segments, avant d'appliquer le regroupement global. Au terme du traitement, chaque locuteur détecté est représenté par une liste de segments temporels indiquant ses instants de parole. Cet outil a fait ses preuves en obtenant de très bons résultats lors de la campagne d'évaluation ESTER2 [Galliano 09].

Sur la figure 1.6, nous avons représenté, superposé au signal audio correspondant, le résultat de la segmentation en locuteurs : 6 locuteurs ont été trouvés, $L = \{loc_1, loc_2, \dots loc_M\}$ avec $M = 6$, et les segments correspondant à chacun sont visualisés.

Pour un locuteur loc_j avec $j = 1 \dots M$, correspond un ensemble S_{loc_j} de N_j segments soit $S_{loc_j} = \{s_{(1,loc_j)}, s_{(2,loc_j)}, \dots, s_{(N_j,loc_j)}\}$. Chaque segment est repéré temporellement par ses instants de début et de fin.

À partir de ces ensembles de segments $\{S_{loc_j}, j = 1 \dots M\}$, nous réalisons la détection des zones d'interaction entre des paires de locuteurs.

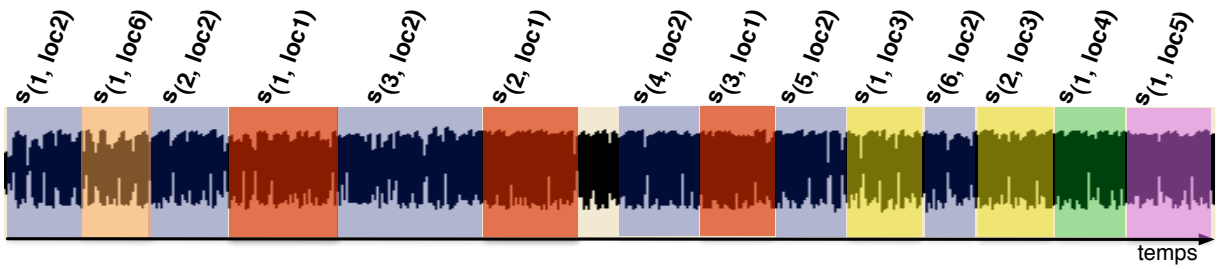


FIGURE 1.6 – Les segments de parole des locuteurs obtenus en sortie du module de SRL.

1.3.3 Détection des zones d'interaction et calcul du niveau d'interactivité

Quand le contenu du message prononcé est disponible, deux tours de parole peuvent être suffisants pour détecter une interaction orale entre deux individus. Un échange du type « Question du locuteur 1 - Réponse du locuteur 2 » tient en deux tours de parole consécutifs. Dans notre étude nous travaillons sans informations linguistiques, ni lexicales, uniquement à partir des segments de parole des locuteurs. Dans ce cas, un échange basé sur seulement deux segments n'est pas assez discriminant. Nous posons l'hypothèse qu'au minimum 3 segments sont nécessaires pour détecter une zone d'interaction entre deux intervenants. Ces trois segments forment un motif que nous nommons une « unité d'interaction ».

1.3.3.1 Définition de l'unité d'interaction

L'unité d'interaction est une série de 3 segments de parole formant une alternance entre deux locuteurs. L'unité d'interaction entre un couple de locuteurs $\{loc_j - loc_{j'}\}, j \neq j'$, est définie par :

$$s(i, loc_j) \; - \; s(k, loc_{j'}) \; - \; s(i+1, loc_j).$$

FIGURE 1.7 – Exemple d'unité d'interaction du couple $\{loc_j - loc_{j'}\}$.

Pour assurer une cohérence à l'unité d'interaction d'un couple de locuteurs, il est nécessaire que les segments de parole consécutifs ne soient pas séparés par un autre locuteur, ou par une zone sans parole de durée d supérieure à un seuil. L'idéal est d'avoir des segments proches les uns aux autres. Dans la pratique, nous tolérons un seuil d égal à une seconde. Par combinaison des unités d'interaction nous générons les zones d'interaction (cf. figure 1.3.3.1).

1.3.3.2 Recherche des zones d'interaction

Les zones d'interaction sont recherchées pour chaque couple de locuteurs $\{loc_j - loc_{j'}\}$ avec $j = 1 \ldots M$, $j' = 1 \ldots M$ et $j \neq j'$ pris parmis l'ensemble des locuteurs extraits par la SRL.

Sur la figure 1.8 nous représentons les Unités d'Interaction (U.I.) découvertes pour trois couples de locuteurs :

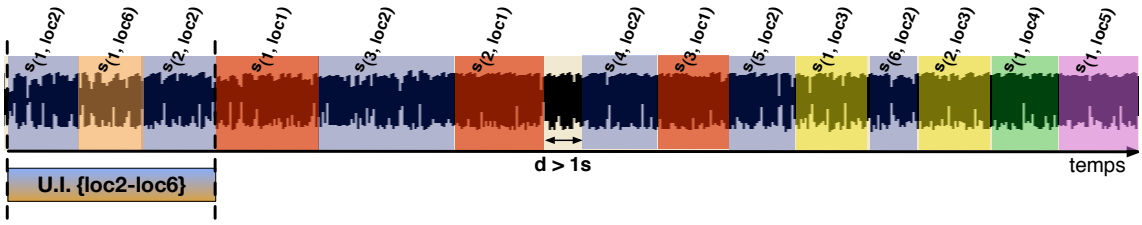
- 1 segment pour le couple $\{loc_2 - loc_6\}$ sur la sous-figure 1.8(a),
- 2 segments pour le couple $\{loc_2 - loc_3\}$ sur la sous-figure 1.8(b),
- 2 segments pour le couple $\{loc_1 - loc_2\}$ sur la sous-figure 1.8(c).

Les autres couples de locuteurs ne présentent pas d'unités d'interaction.

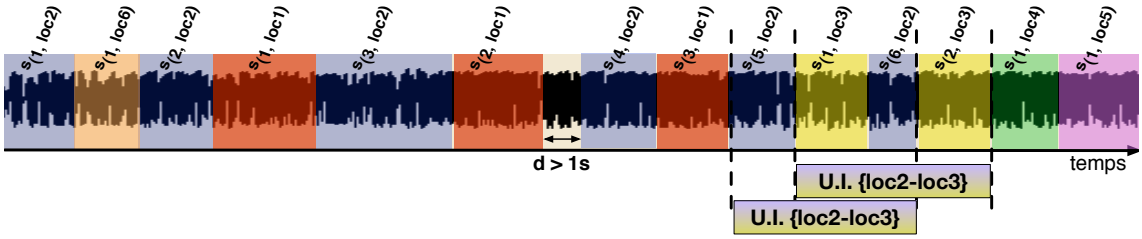
Sur la sous-figure 1.8(c), nous pouvons également voir que deux unités d'interaction du couple de locuteurs $\{loc_1 - loc_2\}$ ont été rejetées car elles contiennent des segments séparés par une durée supérieure au seuil d.

Les Zones d'Interaction (Z.I.) sont obtenues par l'union des unités d'interaction qui se superposent pour un couple de locuteurs donné.

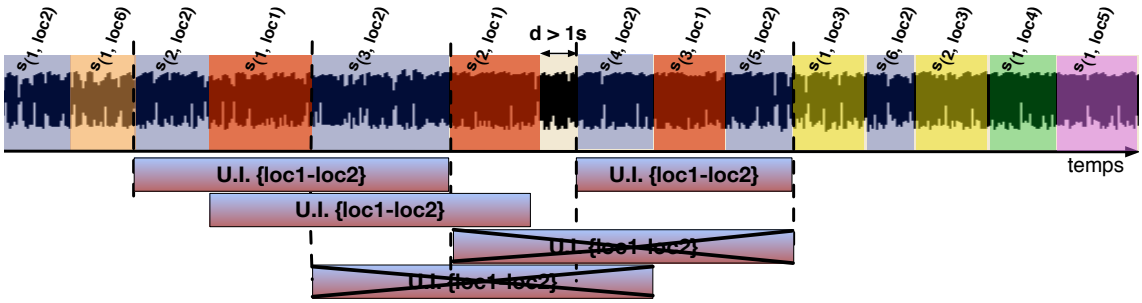
Le niveau d'interactivité est une mesure permettant d'évaluer le potentiel conversationnel d'une zone d'interaction. Plus le nombre d'alternances de tours de parole est important, plus la zone d'interaction est susceptible de contenir de la parole conversationnelle et plus le niveau d'interactivité de la zone est élevé. Le niveau d'interactivité correspond au nombre d'unités d'interaction constitutives d'une zone d'interaction. Le niveau 1 correspond au cas le plus simple où une zone d'interaction correspond à une unité d'interaction. Tout ajout de segment dans la zone d'interaction incrémente le niveau d'interactivité.



(a) Unité d'interaction du couple de locuteurs {loc2 - loc6}



(b) Unités d'interaction du couple de locuteurs {loc2 - loc3}



(c) Unités d'interaction du couple de locuteurs {loc1 - loc2}

FIGURE 1.8 – Exemples d'unités d'interaction.

En reprenant l'exemple précédent (cf. figure 1.8), sur la figure 1.9 nous avons représenté les zones d'interaction repérées par leur position dans le document ainsi que leur niveau d'interactivité :

- les locuteurs loc_2 et loc_6 interagissent une fois dans une zone d'interaction de niveau 1,
- les locuteurs du couple $\{loc_2 - loc_3\}$ apparaissent sur une zone d'interaction de niveau 2,
- le couple $\{loc_1 - loc_2\}$ se retrouve dans deux zones d'interaction, une première de niveau 2, et une seconde de niveau 1.

Certains locuteurs peuvent être impliqués dans des interactions avec plusieurs locuteurs différents, c'est pourquoi nous définissons « l'ensemble des zones d'interaction » d'un locuteur.

L'ensemble des zones d'interaction d'un locuteur $\mathbf{loc_j}$ correspond à l'ensemble ZI_{loc_j} des zones d'interaction dans lesquelles le locuteur loc_j intervient. Dans l'exemple de la figure 1.9, le locuteur loc_2 intervient dans 4 zones d'interaction, loc_1 intervient dans 2 zones d'interaction et les locuteurs loc_3 et loc_6 interviennent dans une seule zone d'interaction. Pour tous les autres locuteurs, l'ensemble des zones d'interaction est vide.

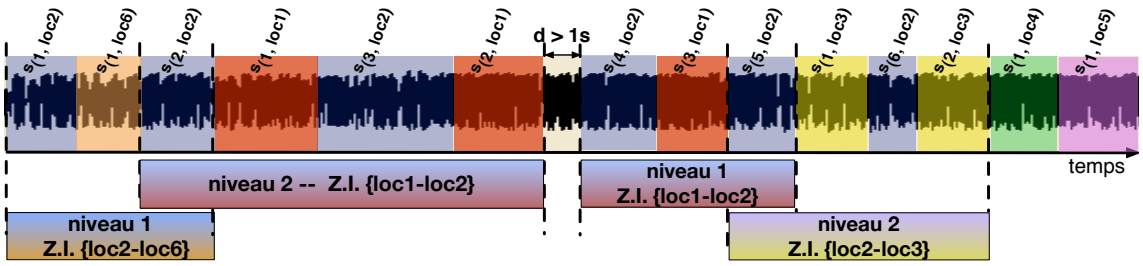


FIGURE 1.9 – Niveaux d'interactivité des Z.I., sur le même exemple que précédemment.

Il est à noter que des zones d'interaction appartenant à des couples de locuteurs différents peuvent se recouvrir sur la longueur d'un segment. Il n'est pas possible de déterminer si le segment commun aux deux zones d'interaction doit être attribué préférentiellement à l'une ou à l'autre des zones, ou aux deux. Par exemple, sur la figure 1.9, le segment $s(2, loc_2)$ appartient à deux zones d'interaction.

Nous proposons sur la figure 1.10 une représentation temporelle des zones d'interaction détectées par cette méthode sur un épisode du débat de société *Le Téléphone Sonne* diffusé sur la station de radio France Inter. Cette représentation a été établie à partir d'une segmentation en locuteurs manuelle. Chaque segment (rouge ou jaune) représente une zone d'interaction. La largeur de ces segments permet de visualiser la durée de l'interaction. La hauteur représente le niveau d'interactivité de la zone d'interaction. Les segments de couleur jaune indiquent que leur niveau est égal à 1, les segments rouges ont un niveau d'interactivité supérieur à 1.

Cette représentation (cf. figure 1.10) met en évidence des zones d'interaction de formats et de durées variés. Nous voulons identifier et caractériser les zones les plus caractéristiques d'une interaction orale. Dans la suite nous proposons d'analyser les zones d'interaction à travers une étude focalisée sur les interventions des locuteurs.

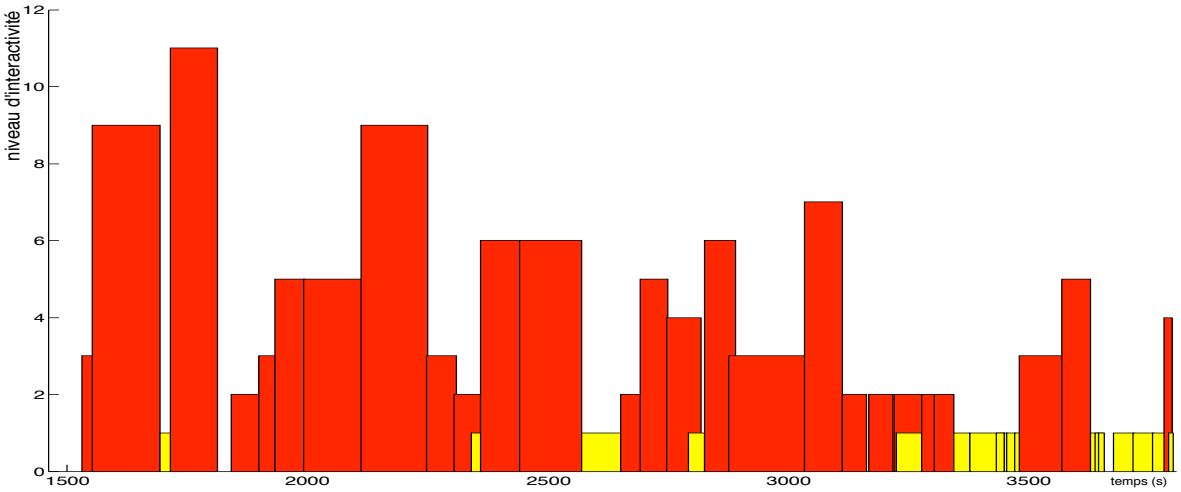


FIGURE 1.10 – Zones d'interaction et niveau d'interactivité pour le débat de société *Le Téléphone Sonne*.

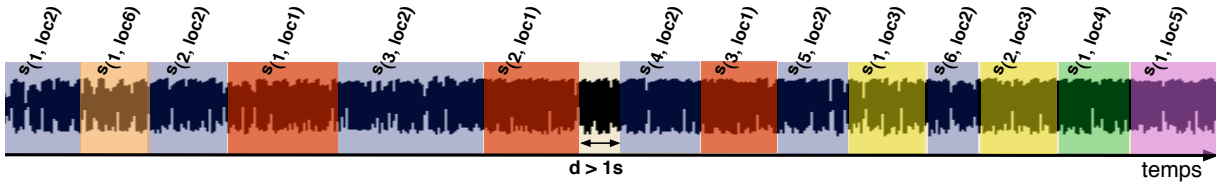
1.3.4 Caractérisation des zones d'interaction

Dans cette étude, nous introduisons un ensemble de descripteurs temporels dans le but d'étudier les interventions des locuteurs. Ces travaux, présentés dans [Bigot 08a, Bigot 08b, Bigot 08c], permettent de caractériser les zones d'interaction.

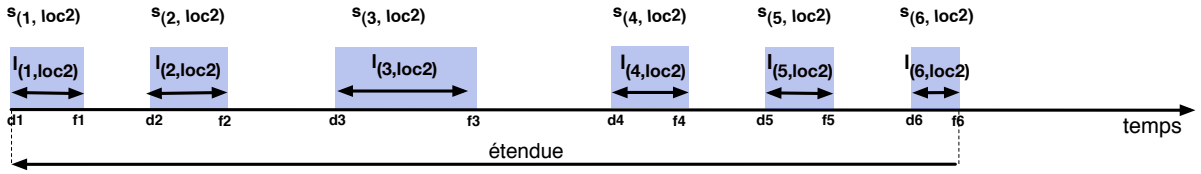
Les interventions des locuteurs d'un même document peuvent être très différentes. Nous nous appuyons sur l'exemple de la figure 1.11 pour illustrer nos remarques :

- certains locuteurs (comme le locuteur loc_1) interviennent à plusieurs reprises.
- d'autres locuteurs (comme loc_4 et loc_5) n'interviennent qu'une seule fois.
- des intervenants apparaissent au début du document et n'apparaissent plus ensuite (comme loc_6). D'autres intervenants (comme loc_2) interviennent sur tout le document.
- certains locuteurs participent à des zones d'interaction (comme loc_1) quand d'autres locuteurs n'interagissent avec personne (comme loc_4).

Nous voulons mesurer ces différences (et points communs) pour caractériser le locuteur et enrichir les zones d'interaction.



(a) Segmentation en locuteurs complète



(b) Segments du locuteur loc_2

FIGURE 1.11 – Exemple de segmentation en locuteurs.

Avant de présenter les descripteurs, nous établissons une catégorie de locuteurs : la catégorie des « locuteurs ponctuels ».

Les locuteurs ponctuels sont les intervenants qui n'apparaissent que sur un seul segment. C'est le cas par exemple du locuteur loc_6 sur la figure 1.11(a). Les locuteurs ponctuels ne peuvent participer qu'à une seule zone d'interaction de niveau d'interactivité égale à 1. C'est une information qui nous permet de considérer ces locuteurs comme peu susceptibles de participer à une séquence conversationnelle.

Nous faisons appel à deux ensembles de descripteurs :

- un ensemble caractérisant l'intervention du locuteur dans sa globalité,
- un ensemble visant à prendre en compte une éventuelle évolution temporelle.

1.3.4.1 Les descripteurs globaux

Pour chaque locuteur nous définissons l'activité globale, l'étendue et la contribution à des zones d'interaction. Ces descripteurs ne dépendent pas de la durée du document.

Pour donner précisément ces termes, nous rappelons et complétons les définitions suivantes :

- N_j représente le nombre de segments du locuteur loc_j.
- Pour chaque locuteur loc_j, nous disposons de l'ensemble S_{loc_j} de ses segments :

$$S_{loc_j} = \{s_{(1,loc_j)}, s_{(2,loc_j)}, \ldots, s_{(N_j,loc_j)}\}$$

  Nous représentons sur la sous-figure 1.11(b) les segments du locuteur loc_2 extraits de la segmentation de la sous-figure 1.11(a).

- $l(k, loc_j)$, $d(k, loc_j)$ et $f(k, loc_j)$ sont respectivement la longueur, le début et la fin du $k^{ième}$ segment $s(k, loc_j)$ du locuteur loc_j.

- **L'activité globale A du locuteur** est le temps de parole cumulé de l'intervenant, ce qui donne :

$$A_{loc_j} = \sum_{k=1}^{N_j} l_{(k,loc_j)}$$

- **L'étendue E du locuteur** mesure sa durée d'apparition. L'étendue d'un locuteur est la durée qui sépare le début de sa première intervention et la fin de sa dernière intervention. Une illustration de cette définition est indiquée sur la figure 1.11(b).
  L'étendue du locuteur loc_j se calcule par :

$$E_{loc_j} = f_{(N_j,loc_j)} - d_{(1,loc_j)}$$

- **la contribution C du locuteur à des zones d'interaction** mesure la proportion de l'activité globale **A** incluse dans des zones d'interaction. Soit $(S, ZI)_{loc_j}$ l'intersection entre l'ensemble des segments S_{loc_j} du locuteur loc_j et l'ensemble des zones d'interaction de ce locuteur ZI_{loc_j}. Ce descripteur est la durée cumulée des segments appartenant à $(S, ZI)_{loc_j}$ divisée par l'activité globale du locuteur A_{loc_j} :

$$C_{loc_j} = \frac{\text{durée}(S_{loc_j} \bigcap ZI_{loc_j})}{A_{loc_j}}$$

1.3.4.2 Les descripteurs locaux

Le contenu d'un document audiovisuel évolue autour de l'apparition et du départ des intervenants. Par exemple, le locuteur loc_6 de la figure 1.12 intervient au début du document uniquement, le locuteur loc_5 au contraire apparaît à la fin du document. Pour quantifier dans quelles parties du document se répartissent les interventions des locuteurs, nous proposons de calculer l'activité locale du locuteur, les activités locales sont ensuite regroupées au sein d'un vecteur de répartition de l'activité.

Ces descripteurs viennent compléter les descripteurs précédents en intégrant une dimension dynamique au calcul de l'activité de parole d'un locuteur. Nous posons W un ensemble de U fenêtres temporelles d'analyse réparties tout au long du document, adjacentes et d'intersection vide 2 à 2, soit $W = \{w_1, w_2, \ldots, w_U\}$. Sur la figure 1.12 nous représentons le cas où $U = 3$.

- **L'activité locale** $a_{(w_i,loc_j)}$ sur la fenêtre w_i d'un locuteur loc_j est la durée relative de l'intersection entre l'ensemble de segments du locuteur, S_{loc_j}, et la fenêtre d'analyse w_i :

$$a_{(w_i,loc_j)} = \text{durée}(S_{loc_j} \bigcap w_i)$$

  L'activité locale sur une fenêtre représente une fraction de l'activité globale d'un locuteur.

- **Le vecteur de répartition de l'activité**, noté $\mathbf{a}_{(W,loc_j)}$ du locuteur loc_j est formé des activités locales calculées pour chaque fenêtre d'analyse de W de telle sorte que :

$$\mathbf{a}_{(W,loc_j)} = [a_{(w_1,loc_j)} \quad a_{(w_2,loc_j)} \quad \ldots \quad a_{(w_U,loc_j)}]$$

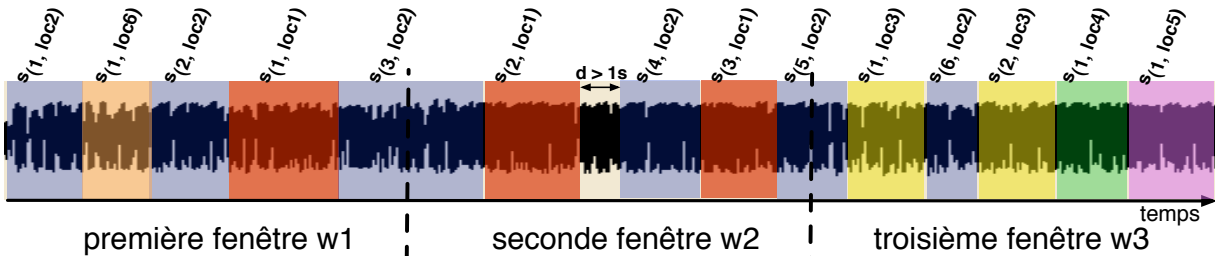


FIGURE 1.12 – Découpage d'une segmentation en locuteurs en trois sections de même durée.

Pour l'exemple proposé sur la figure 1.12, où $U = 3$, le vecteur de répartition de l'activité de chaque locuteur est calculé et reporté dans la table 1.1.

TABLE 1.1 – Vecteurs de répartition de l'activité locale pour l'exemple de la figure 1.12.

$$\begin{array}{lcccc}
\mathbf{a}_{(W,loc_1)} = & [& 0.4\times A_{loc_1} & 0.6\times A_{loc_1} & 0\times A_{loc_1}] \\
\mathbf{a}_{(W,loc_2)} = & [& 0.45\times A_{loc_2} & 0.4\times A_{loc_2} & 0.15\times A_{loc_2}] \\
\mathbf{a}_{(W,loc_3)} = & [& 0\times A_{loc_3} & 0\times A_{loc_3} & 1\times A_{loc_3}] \\
\mathbf{a}_{(W,loc_4)} = & [& 0\times A_{loc_4} & 0\times A_{loc_4} & 1\times A_{loc_4}] \\
\mathbf{a}_{(W,loc_5)} = & [& 0\times A_{loc_5} & 0\times A_{loc_5} & 1\times A_{loc_5}] \\
\mathbf{a}_{(W,loc_6)} = & [& 1\times A_{loc_6} & 0\times A_{loc_6} & 0\times A_{loc_6}]
\end{array}$$

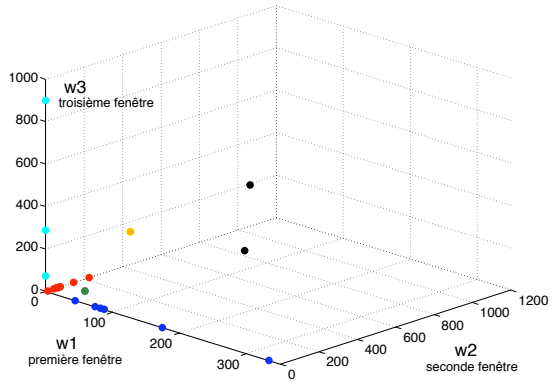
1.3.4.3 Mise en évidence de l'activité locale suivant le type de programme

La figure 1.13 rassemble plusieurs exemples de vecteurs de répartition de l'activité calculés avec trois fenêtres d'analyse. Les vecteurs de répartition sont projetés dans un espace à

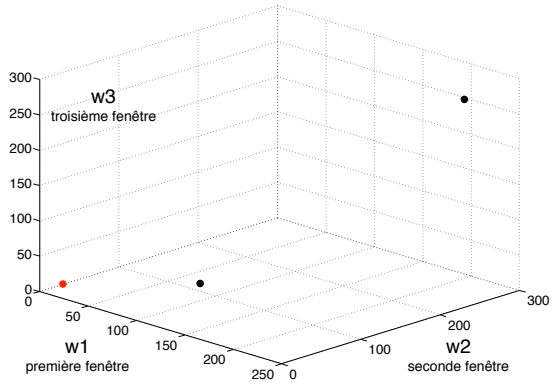
3 dimensions pour les locuteurs de trois émissions radiophoniques différentes. Dans cet espace de représentation, chaque dimension correspond à une fenêtre d'analyse temporelle. La sous-figure 1.13(a) correspond à une émission de type « matinales », la sous-figure 1.13(b) à un magazine culturel et la sous-figure 1.13(c) rassemble les locuteurs d'un débat de société.

Les locuteurs peuvent être rassemblés en 3 catégories :

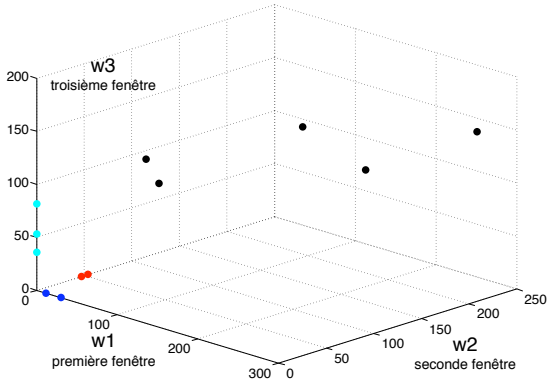
- les locuteurs actifs sur une seule fenêtre d'analyse se rassemblent dans un premier groupe représenté sur le plan par des points de couleur bleue, rouge et cyan,
- les locuteurs actifs sur deux fenêtres d'analyse consécutives correspondent aux points colorés en vert, jaune et magenta,
- les locuteurs qui parlent sur les trois fenêtres sont indiqués en noir sur la figure 1.13.



(a) Émission de type matinales



(b) Magazine culturel



(c) Débat de société

FIGURE 1.13 – Vecteurs de répartition de l'activité locale des locuteurs pour trois émissions radiophoniques (cas $U = 3$).

Sur plusieurs exemples d'émission nous observons un lien entre le type de programme et le nombre de locuteurs de chaque catégorie. La représentation des vecteurs de répartition des locuteurs est une piste à considérer en perspective d'un regroupement par type de programme.

Dans la suite de ce travail, nous étudions également les descripteurs globaux en vue d'une catégorisation des locuteurs.

1.3.4.4 Une première typologie des locuteurs

Dans cette étude, nous cherchons à interpréter la répartition des locuteurs issus de plusieurs documents, dans le plan (activité, étendue). Sur cette représentation dont un exemple est donné sur la figure 1.14 :

- l'axe des abscisses correspond à l'activité globale des locuteurs, normalisée par l'étendue du locuteur le plus présent dans le document,
- l'axe des ordonnées correspond au rapport entre l'étendue du locuteur et l'étendue maximale du document.

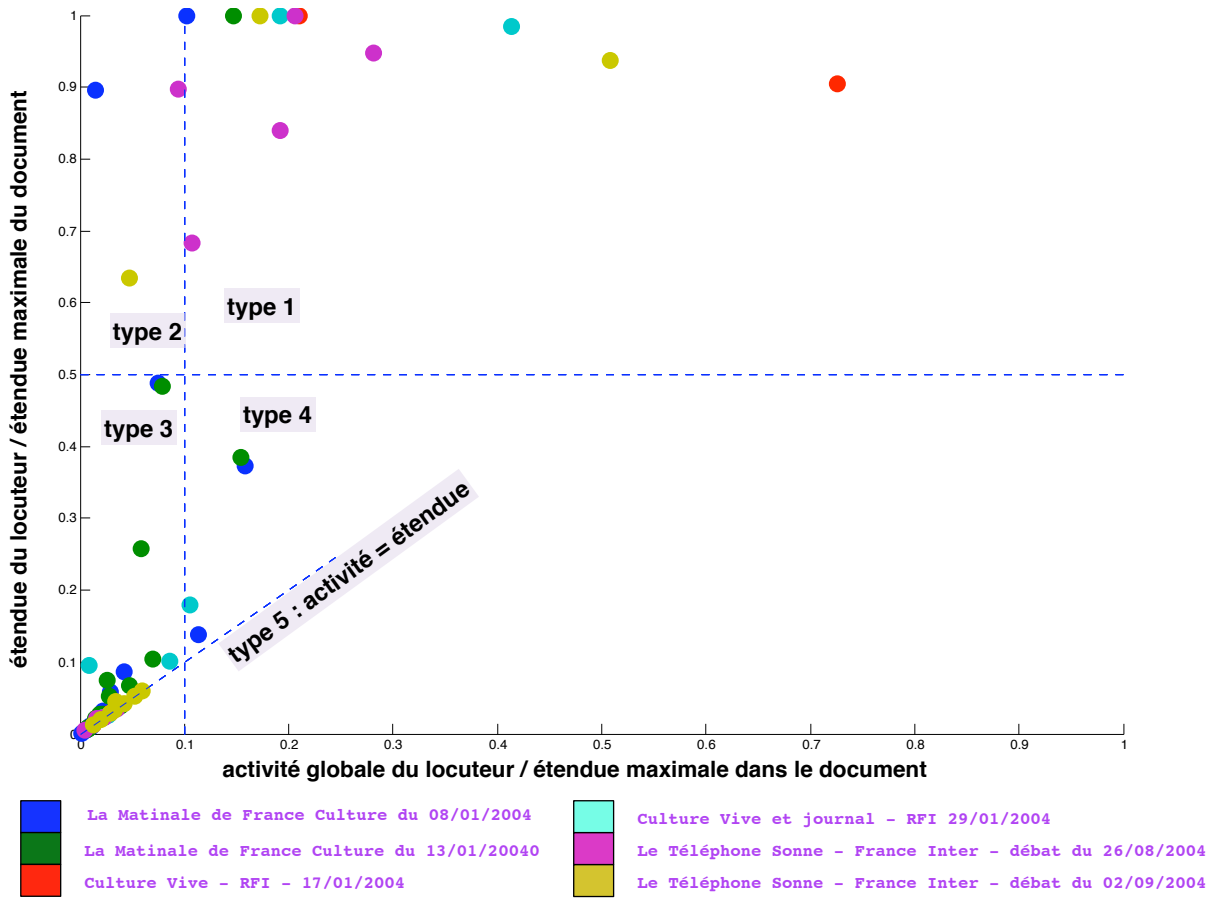


FIGURE 1.14 – Représentation des locuteurs de six documents en fonction de leur activité globale et de leur étendue.

L'exemple de la figure 1.14 rassemble des locuteurs de six émissions : deux émissions matinales, deux magazines, un journal et deux débats de société.

Suite à une étude sur un plus large ensemble de documents, nous proposons de découper le plan (activité, étendue) en 4 quadrants que nous supposons correspondre à 4 types d'intervenant, auxquels s'ajoute une cinquième catégorie correspondant aux locuteurs situés sur la première diagonale.

Nous pouvons décrire chaque type de locuteur :

- les locuteurs de type 1 ont une activité et une étendue importante. Il serait raisonnable de trouver des présentateurs, des invités principaux et plus généralement des locuteurs importants car très présents dans le document.
- le type 2 correspond à des locuteurs peu actifs mais très étendus. Ce sont des locuteurs qui apparaissent sur des segments courts et distants les uns des autres. Ce type de locuteurs correspond typiquement aux interventions d'un annonceur ou à une séquence telle qu'un flash d'information court et revenant régulièrement.
- les locuteurs de type 3 sont les plus nombreux. Ces intervenants sont peu actifs et peu étendus. Ils font généralement des interventions localisées à une partie du document. Les chroniqueurs, les journalistes ou les personnes participant à une interview relèvent de cette catégorie.
- les locuteurs de type 4 ont une activité importante et une étendue plus faible. Ce type est caractéristique de locuteurs parlant beaucoup mais de manière très dense et très localisée. C'est le cas couramment des personnes interviewées durant un bulletin d'information.
- la dernière catégorie correspond aux locuteurs ponctuels situés sur la droite de pente unité. Les locuteurs ponctuels ne sont actifs que sur un seul segment, de ce fait leur activité est égale à leur étendue. Ce type de locuteur correspond à des envoyés spéciaux et des journalistes dans les journaux. Du fait de leur ponctualité, ces locuteurs n'interviennent quasiment jamais dans des zones d'interaction.

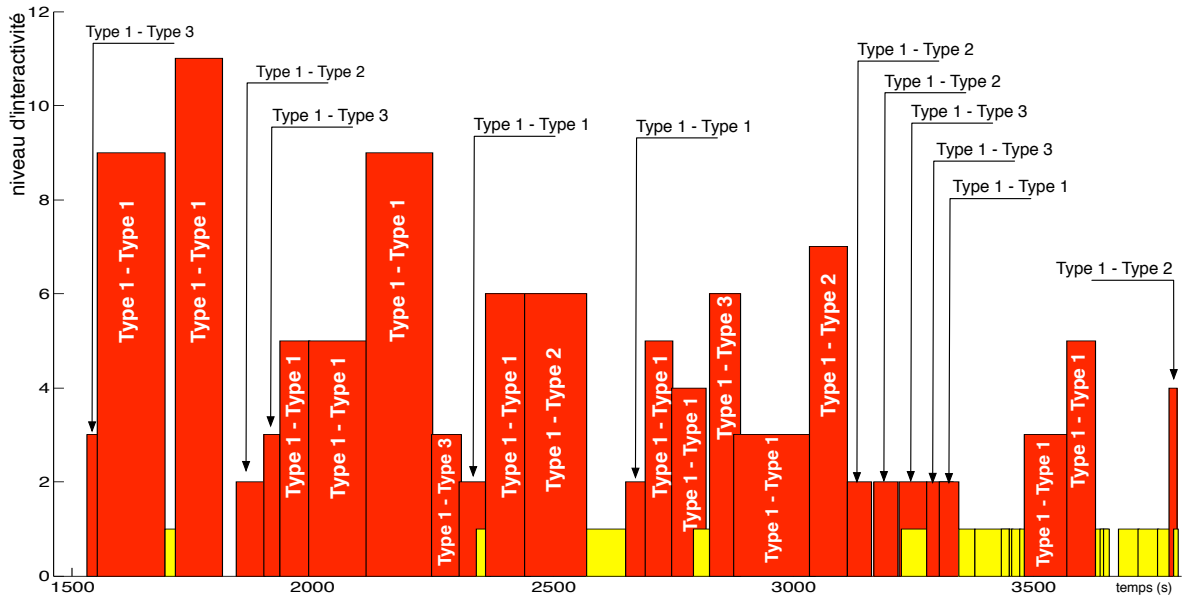


FIGURE 1.15 – Zones d'interaction, niveau d'interactivité et type de locuteur pour le débat de société *Le Téléphone Sonne*.

La figure 1.15 est un exemple de représentation enrichie des zones d'interaction : les informations relatives aux types de locuteurs sont ajoutées aux zones d'interaction (il s'agit de

l'enregistrement *Le Téléphone Sonne*, présenté auparavant, cf. figure 1.10). Le présentateur et les invités principaux, sont de Type 1. Un autre invité principal est Type 2. Sur cette figure, les zones (Type1-Type1), (Type1-Type2) sont des interactions entre le présentateur et un invité, ou entre deux invités.

Durant cette émission des auditeurs téléphonent au standard de l'émission pour poser leurs questions aux invités principaux. Ces locuteurs sont identifiés comme appartenant au Type 3. Ces intervenants interagissent avec le présentateur ou les invités comme l'indiquent plusieurs zones d'interaction de niveaux 3 et 6.

Dans le cadre du projet EPAC, la détection des zones d'interaction (enrichie avec les informations relatives aux types de locuteurs qui y sont impliqués) a été réalisée sur l'ensemble des données du projet. Les résultats ont été fournis sous un format structuré : une description en est faite en annexe B de ce manuscrit.

1.4 Conclusion

Dans ce chapitre, nous avons présenté une étude préliminaire fondée autour de la recherche de zones temporelles d'un document pouvant potentiellement contenir une conversation entre deux locuteurs. Nous avons nommé ces zones : zones d'interaction orale. La méthode que nous avons proposé exploite uniquement la composante audio du flux audiovisuel, à travers uniquement la connaissance *a priori* des tours de parole des intervenants. Elle est également indépendante de la transcription de la parole relative à ces zones et se fonde sur l'analyse de l'organisation temporelle des tours de parole d'un couple de locuteurs. Une première caractérisation des zones d'interaction consiste à évaluer le nombre d'alternances de tours de parole de la zone. Nous nommons cette mesure : le niveau d'interactivité. D'un point de vue de la structuration des documents audiovisuels, les zones d'interaction sont *a priori* des éléments structurants du contenu qui, avec une granularité fine détaillent des événements riches d'enseignement sur le contenu des documents. Nous avons donc souhaité faire évoluer notre compréhension des zones d'interaction.

Le niveau d'interactivité permet des comparer plusieurs zones entre elles, mais nous souhaitons aller plus loin dans la caractérisation des zones d'interaction en étudiant plus particulièrement les caractéristiques propres aux locuteurs impliqués dans les conversations. Nous proposons plusieurs paramètres temporels, globaux et locaux, qui nous permettent d'observer des différences importantes entre les intervenants d'un même document et d'une même zone d'interaction. Finalement nous proposons une catégorisation des locuteurs répartis parmi 5 catégories en fonction de leur étendue et de leur activité, c'est-à-dire de leur temps d'apparition et de leur temps de parole total.

Un lien semble exister entre ces cinq catégories et les rôles réels des intervenants (présentateur, journaliste, invité...). C'est donc tout naturellement que nous poursuivons notre travail de recherche autour de la caractérisation des rôles des intervenants. Nous présentons notre contribution à ce domaine dans le chapitre suivant.

Chapitre 2

Des paramètres pertinents pour la reconnaissance automatique des rôles

L'étude sur la caractérisation des interactions orales entre intervenants dans des documents audiovisuels a mis en évidence cinq catégories d'intervenants se distinguant par l'organisation temporelle de leurs tours de parole. En particulier, nous avons observé que certains locuteurs parlent beaucoup et à de nombreuses reprises, quand d'autres au contraire, font des apparitions très courtes et peu nombreuses. Nous avons mis en relief le lien possible entre ces informations temporelles et le rôle de l'intervenant dans le document. Ces observations fondent les travaux que nous présentons dans ce chapitre, autour de la reconnaissance automatique du rôle de l'intervenant.

Le rôle est un concept sociologique qui peut se définir comme l'ensemble des normes comportementales que doit adopter un individu pour être en accord avec la représentation commune de sa fonction et de sa position sociale [Biddle 86]. Le rôle d'un intervenant a une influence sur son comportement et sur ses interactions avec d'autres individus.

Dans notre étude préliminaire, nous avons mis en relief que dans le contexte d'émissions de radio généralistes, les locuteurs « qui parlent le plus » correspondent souvent à des présentateurs ou à des animateurs. Cette présence importante du présentateur tout au long de l'émission est liée à sa fonction puisque ce locuteur est en charge d'introduire l'émission, de présenter ses invités, de lancer les diverses séquences et de clôturer l'émission. Toutes ces actions font de lui un intervenant central et récurrent. Nous avons dans le même temps observé que les informations temporelles extraites des tours de parole atteignent rapidement leurs limites et ne permettent pas de distinguer aussi clairement d'autres types de rôles. Nous proposons un ensemble de paramètres de natures différentes venant s'ajouter aux paramètres temporels présentés dans le chapitre précédent afin d'approfondir la catégorisation des locuteurs. L'objectif est de trouver des paramètres pertinents dans le cadre d'une reconnaissance automatique du rôle.

Les rôles des intervenants représentent une information importante pour la compréhension du contenu d'un document audiovisuel. Des travaux ont mis en évidence le lien entre les rôles des intervenants et la structure de documents audiovisuels. Les changements de sujets dans les bulletins d'information en particulier sont liés aux instants de changement de rôles comme nous l'indique [Stolcke 99]. De par son rôle central, les interventions du présentateur peuvent permettre de structurer le contenu d'un document comme rapporté dans [Amaral 03] et [Ma 09].

Dans le travail de [Kolluru 07], les rôles des locuteurs sont utilisés pour classer en plusieurs catégories les histoires extraites de bulletins d'information. Les rôles sont utilisés dans le travail de [Weng 07] sur un corpus de films pour structurer les longs métrages en plusieurs histoires parallèles. Les rôles peuvent également être utilisés pour préparer des résumés de documents comme dans [Maskey 03]. La connaissance des rôles peut faciliter la navigation dans un document en permettant d'accéder directement à une intervention particulière.

L'information du rôle peut être une information importante pour les systèmes de transcription automatique de la parole en vue d'une adaptation du lexique ou du modèle de langage. Un journaliste par exemple aura tendance à lire un texte, quand un invité dans une interview aura plutôt tendance à parler de manière moins préparée. Une des fonctions du présentateur est souvent d'introduire les autres intervenants, une étude de la transcription de la parole de ce locuteur central peut permettre d'obtenir les noms des personnes présentes dans l'émission. Ces noms sont alors exploités pour la détection des intervenants nommés [Estève 07].

Ce chapitre s'articule autour de 4 sections. Dans la première partie, nous présentons un état de l'art des méthodes de reconnaissance automatique des rôles des locuteurs. Cette revue nous amène à préciser la définition des rôles donnée dans la section 2.2. La troisième partie est consacrée à la description des paramètres pertinents que nous proposons pour la reconnaissance automatique des rôles des locuteurs. La dernière section est consacrée à une première validation expérimentale de ces paramètres sur un corpus radiophonique.

2.1 État de l'art de la reconnaissance automatique des rôles des locuteurs

La reconnaissance automatique des rôles dans les documents audiovisuels est un sujet de recherche récent, et les travaux publiés sont encore relativement peu nombreux. Les premières méthodes ont été appliquées à des émissions d'information (radio et télévision) et se fondaient essentiellement sur les transcriptions des documents. Le contenu très préparé et la présence de parole lue ont permis à cette approche d'atteindre de bonnes performances en reconnaissance des rôles. Des travaux se sont ensuite orientés vers des enregistrements de groupes de travail (ou meetings). Ces documents, dans lesquels la parole est plutôt de nature conversationnelle, voire très spontanée, ont mis en évidence les limites de l'approche basée sur les transcriptions. De nouvelles contributions, plus récentes, proposent d'exploiter l'organisation temporelle des tours de parole des locuteurs ainsi que les méthodes d'analyse des réseaux sociaux pour détecter le rôle des intervenants.

Nous retiendrons l'existence de deux types d'approches qui structureront la suite de cette section. La première stratégie exploite les transcriptions manuelles ou automatiques de la parole à travers l'extraction de paramètres lexicaux. La seconde se base sur une analyse de l'organisation des tours de parole des locuteurs issus d'une segmentation et d'un regroupement en locuteurs (manuels ou automatiques).

2.1.1 Détection des rôles basée sur l'analyse des transcriptions

L'analyse du vocabulaire utilisé par les intervenants est la première piste suivie pour reconnaître un rôle :

- Les phrases clés d'introduction ou de conclusion souvent redondantes, sont de véritables signatures de la fonction de présentateur ou animateur ;
- Les différences entre la parole préparée des journalistes et des présentateurs et la nature plus spontanée des interventions des invités (reprise, hésitation...) s'imposent tout autant.

C'est pourquoi les premières études de reconnaissance de rôles se sont faites à partir des résultats de transcription automatique du discours.

Le travail de Barzilay [Barzilay 00] est à notre connaissance une des premières contributions de la littérature rapportant les performances d'une reconnaissance automatique en rôles. Le corpus se compose de bulletins d'informations. Trois rôles sont recherchés – *présentateur*, *journaliste* et *invité* – dans l'objectif de produire des résumés structurels des documents. La reconnaissance automatique se fonde sur le temps de parole des locuteurs ainsi que sur la recherche dans les transcriptions :

- de regroupements lexicaux fréquents pouvant être caractéristiques des rôles ;
- de groupes de mots durant lesquelles les locuteurs se présentent, ou présentent explicitement d'autres personnes.

Aux paramètres du segment courant sont joints les mêmes paramètres extraits sur les n segments précédents. L'évaluation est menée sur 35 enregistrements (soit 17 heures) d'une même émission de radio, c'est-à-dire avec un contenu et une structure très similaire. Le taux de reconnaissance correcte est de 80% de segments bien étiquetés.

Les travaux de Canseco [Canseco 05] concernent la segmentation et le suivi des locuteurs. L'objectif est dans ce cas d'associer le nom d'un locuteur avec un segment de parole. Les auteurs développent une approche qui met à profit la connaissance des rôles des locuteurs. Quatre rôles sont étudiés : *présentateur*, *journaliste*, *invité* et *annonceur*. Canseco cherche tout d'abord les séquences lexicales les plus fréquentes relatives à chacun des rôles cités. Il découvre 11 locutions caractéristiques des présentateurs, 20 locutions pour les journalistes, 8 locutions pour les invités et 9 locutions pour les annonceurs. Une dernière étape lie les rôles aux noms réels des locuteurs en utilisant des formes lexicales généralisées, déterminant ainsi si les noms cités se rapportent au locuteur courant, au locuteur suivant ou au précédent. Des expériences sont réalisées sur près de 150 heures de données de bulletins d'information anglophones.

En 2006 Liu [Liu 06] propose une méthode fondée sur la combinaison de deux approches : les Modèles de Markov Cachés (MMC) et le maximum d'entropie.

- La topologie des MMC considère un rôle comme étant un état du modèle. Les observations sont les séquences de mots prononcés par un locuteur durant un tour de parole.
- Le classifieur à maximisation d'entropie utilise seulement le premier et le dernier mot d'un tour de parole.

L'attribution des rôles des locuteurs se fait parmi 3 catégories : *présentateur*, *journaliste* et *autre*. Des transcriptions de parole manuelles sont utilisées pour apprendre les modèles de chaque

rôle. Le corpus se compose de 170 heures de données audio (336 bulletins d'information de plusieurs sources en mandarin). Chacune de ces approches atteint environ 77% de classification correcte en rôles. Les deux approches sont combinées, ce qui permet d'atteindre 80% de taux de reconnaissance correcte.

La dernière contribution rapportée dans cette partie concerne la reconnaissance des rôles dans des enregistrements de groupes de travail. Cette étude menée par Banerjee [Banerjee 06] en 2006 exploite un corpus de réunions professionnelles simulées dans lesquelles 3 ou 4 personnes jouent des rôles tels que *responsable de réunion*, *expert en acquisition de logiciel* et *expert en logistique*. La reconnaissance des rôles s'appuie sur l'extraction des mots les plus souvent prononcés par chacun des trois rôles : 180 mot-clefs sont identifiés dans les résultats de transcription manuelles. Pour chaque participant, 4 paramètres sont calculés sur une fenêtre d'analyse du signal de parole :

- la proportion du temps de parole d'un participant par rapport au temps de parole total,
- le rang du participant en terme de temps de parole,
- le rang du participant en nombre d'utilisation des mots-clefs,
- le ratio entre le nombre d'utilisations d'un mot-clef par le participant et le nombre total d'utilisations de ce mot.

La reconnaissance atteint 83% de rôles bien attribués.

Ces quatre études présentent globalement de bonnes performances, toutefois ils nécessitent une transcription de la parole de bonne qualité. A l'heure actuelle, les taux d'erreurs sur les contenus conversationnels limitent l'applicabilité de ce type de méthodes : le taux d'erreur-mot obtenu en moyenne par tous les participants de la tâche de transcription lors de la campagne d'évaluation ESTER [Galliano 05] est de 35%. Ce taux d'erreur est descendu à 20% lors de la récente campagne ESTER2 [Galliano 09])

Les méthodes que nous allons présenter dans la suite n'ont pas cette limitation car elles n'utilisent pas le contenu des messages transmis par les intervenants. Elles s'appuient principalement sur l'analyse de l'organisation temporelle des tours de parole des locuteurs rendus disponibles par l'utilisation de méthodes de segmentation et de regroupement en locuteurs (SRL). Cette segmentation est plus robuste comme l'indique le taux d'erreur moyen (12%) obtenu par les participants à cette tâche lors de la campagne ESTER2.

2.1.2 Détection des rôles basée sur les résultats de segmentation et regroupement en locuteurs

En 2007, Vinciarelli [Vinciarelli 07] propose une approche pour la reconnaissance des rôles appliquée à des résultats de SRL. Six rôles sont reconnus en accord avec le contenu du corpus composé de 96 enregistrements de la même émission d'une durée d'environ 12 minutes chacun : *présentateur*, *second présentateur*, *invité*, *participant d'interview*, *annonceur des titres* et *annonceur météo*. L'originalité de son travail repose dans la prise en compte des interactions entre intervenants via l'analyse d'un réseau social, combinée avec une modélisation statistique de la distribution des longueurs de segments de parole de chaque rôle. Les rôles des locuteurs d'un même document sont reconnus successivement en tirant partie des deux types d'information. Les performances atteignent 85% de bonne reconnaissance par rapport à la durée totale traitée.

Cette proposition se retrouve dans les travaux de Salamin [Salamin 09]. Également dans le contexte de l'analyse des réseaux sociaux, ce travail exploite un outil appelé un réseau d'affiliations sociales. Ce réseau permet de quantifier, tout au long du document sur des fenêtres temporelles d'analyse, les interactions entre les locuteurs. Sur chaque fenêtre d'analyse, sont calculés des vecteurs de descripteurs liés à l'interaction entre intervenants. Un modèle statistique est finalement appliqué aux vecteurs de paramètres dans le but d'assigner un rôle à chaque locuteur. Les évaluations sont menées sur le corpus de bulletins d'information utilisé dans [Vinciarelli 07], auquel s'ajoute un corpus de 27 heures d'émissions de société et un corpus de 45 heures de réunions de travail. Le temps de parole correctement étiqueté en rôle atteint 80% sur les deux premiers corpus, mais les performances s'effondrent à 45% sur le corpus de réunions de travail. Sur ce dernier corpus, la segmentation en locuteurs est quasiment parfaite, la pureté de la segmentation en locuteurs est indiquée égale à 99% par les auteurs. Par conséquent, les erreurs de reconnaissance en rôle ne peuvent pas être attribuées aux erreurs de SRL. Ce résultat démontre les limites d'une approche uniquement basée sur une modélisation de l'organisation temporelle des tours de parole pour la reconnaissance des rôles des locuteurs dans des documents peu structurés tels que des réunions de travail.

Nos travaux s'inscrivent dans la lignée des deux dernières méthodes présentées [Bigot 10a, Bigot 10b, Bigot 10c, Bigot 10d]. Ils sont basés sur l'extraction de paramètres bas-niveau de natures variées disponibles à partir d'une SRL, indépendamment de la structure du document. Ils sont développés dans la suite de ce chapitre. Mais avant cela, et à la lumière de ces études, il nous semble nécessaire de préciser la définition de chaque rôle et d'étendre le nombre de catégories.

2.2 Définition des rôles ; la notion d'intervenant « ponctuel »

Les rôles recherchés dans la quasi-totalité des travaux de l'état de l'art sont *présentateur, journaliste, autre (ou invité)*. Cependant, les observations faites lors de l'étude préliminaire présentée dans le chapitre II nous conduisent à prendre en compte la notion de « locuteur ponctuel ». Un locuteur ponctuel est un locuteur qui ne compte qu'une seule intervention (c'est à dire un seul segment de parole). Ce type d'intervenant ne sera pas amené à participer à des séquences conversationnelles avec d'autres locuteurs. Grâce à cette particularité, nous pouvons affiner les définitions des catégories de rôles. Ainsi nous proposons d'étendre la liste des rôles recherchés à cinq rôles en nous appuyant sur la notion de ponctualité : ces rôles sont ***présentateur***, ***journaliste ponctuel***, ***journaliste non ponctuel***, ***autre ponctuel***, ***autre non ponctuel***. Ces rôles sont suffisamment génériques pour correspondre à des corpus composés d'enregistrements d'émissions variées et permettre des comparaisons avec la plupart des performances rapportées dans la littérature. Nous décrivons maintenant chacun de ces rôles.

Présentateur Cet intervenant est en charge de présenter ou d'animer une émission. Il assure généralement l'introduction et la conclusion de l'émission. Il introduit les autres intervenants et lance les séquences du document. Le présentateur peut également interviewer un invité. Les interventions du présentateur sont généralement préparées, particulièrement dans des bulletins

d'information. Ce locuteur a une place centrale et possède souvent un temps de parole important sur un grand nombre de tours de parole répartis tout au long de l'émission.

Journaliste non ponctuel Cette catégorie de rôle rassemble des professionnels intervenant sur plusieurs tours de parole. Ces locuteurs peuvent correspondre plus particulièrement à des chroniqueurs, des intervieweurs et des reporters. Les interventions de ces locuteurs sont souvent préparées et peuvent correspondre à des chroniques avec des tours de parole assez longs ou des entretiens durant lesquels le journaliste ne monopolise pas le temps de parole et laisse à son invité le temps de s'exprimer.

Journaliste ponctuel Cas particulier du journaliste, le journaliste ponctuel n'intervient que sur un seul de tour de parole. Dans cette catégorie de rôle nous allons trouver entre autres, des envoyés spéciaux, des correspondants à l'étranger ou en province, des chroniqueurs de de la bourse ou de la météo. Ces locuteurs n'interagissent avec aucun autre intervenant. Les interventions de ces locuteurs sont souvent très préparées, voire lues.

Les deux dernières catégories correspondent à des intervenants qui ne sont ni présentateur, ni journaliste. Ces classes vont ainsi rassembler une grande variété de locuteurs et il est difficile d'en faire une liste exhaustive. Il s'en suit que leurs interventions peuvent contenir un texte lu ou au contraire une intervention très spontanée.

Autre non ponctuel Cette classe correspond plus particulièrement aux invités d'une émission ou d'une interview. Ces locuteurs peuvent être également des anonymes lorsqu'il s'agit d'auditeurs ou de téléspectateurs intervenant au téléphone pour poser des questions à une personnalité ou à un expert. Un locuteur de cette catégorie peut parler beaucoup dans le document dont il est l'invité principal, mais son intervention peut également se réduire à quelques minutes pour une interview dans le cadre d'un journal télévisé par exemple. Le type de parole utilisé pourra aller de préparée à spontanée.

Autre ponctuel Cette catégorie correspond à des locuteurs apparaissant dans des enregistrements diffusés souvent en différé comme des extraits de conférences de presse ou d'interviews. Il peut s'agir également d'extraits de pièces de théâtre, de films ou d'archives historiques. Ces locuteurs ne participent pas explicitement à des longues séquences conversationnelles, mais leurs interventions peuvent en être extraites. Ils apparaissent au cours d'un seul tour de parole et la durée de l'intervention peut être très variée, de même que la qualité de l'enregistrement.

La répartition temporelle de l'activité de parole mais également la manière de parler des locuteurs interviennent dans les différences qui existent entre ces rôles. Nous avons donc essayé de capter ces informations à travers l'extraction de jeux de paramètres bas-niveau disponibles dans l'analyse des interventions de chaque intervenant. Ces jeux de paramètres sont présentés dans la section suivante.

2.3 Extraction de paramètres temporels, acoustiques et prosodiques

2.3.1 Contribution à la reconnaissance du rôle

En écoutant des programmes radiophoniques ou télévisuels dans une langue étrangère que nous ne connaissons pas, en peu de temps nous sommes capables d'identifier les rôles des intervenants ou de les classer dans des catégories différentes. Ce constat met en avant que nous pouvons identifier des comportements et des caractéristiques des locuteurs sans avoir connaissance du message prononcé.

Dans la mesure où nous pouvons extraire intuitivement certaines spécificités des intervenants nous souhaitons tirer profit des méthodes de reconnaissance des formes et des algorithmes de classification pour exploiter certaines caractéristiques en vue de développer un système de reconnaissance automatique des rôles des intervenants.

Dans ce sens nous fondons notre approche sur l'extraction de trois catégories de paramètres bas-niveau calculés pour chaque locuteur :

- des paramètres temporels, à travers lesquels nous souhaitons capter des informations sur la répartition des interventions du locuteur au cours de temps.
- des paramètres acoustiques permettant de caractériser l'adaptation du locuteur à l'environnement (bruyant ou calme) dans lequel il intervient.
- les paramètres prosodiques puisque le débit de parole et l'intonation d'un locuteur peuvent varier en fonction de son aptitude à parler en public ou du niveau de préparation de son intervention.

Notre proposition repose également sur l'hypothèse que le rôle d'un locuteur reste le même dans un document. Un unique rôle est attribué à un intervenant unique dans un document unique. La figure 2.1 rassemble les étapes de l'extraction des paramètres.

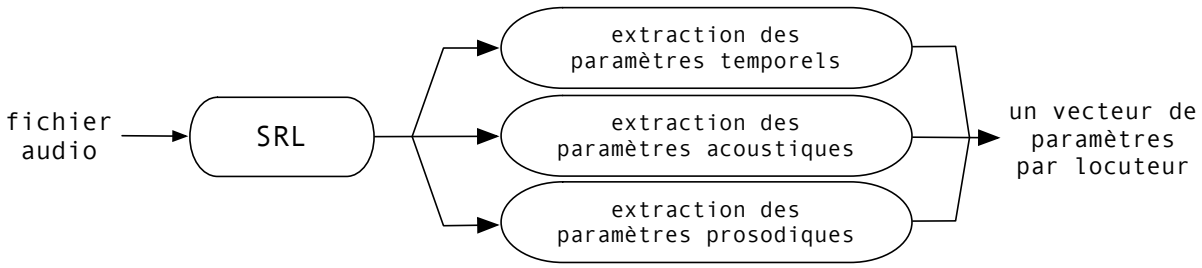


FIGURE 2.1 – Méthode d'extraction des paramètres « bas-niveau ».

Afin d'identifier les instants de parole de chaque locuteur, le document audio est tout d'abord traité par une méthode automatique de segmentation et de regroupement en locuteurs (SRL, c.f. 2.3.2). Dans un second temps, pour chaque locuteur détecté nous calculons l'ensemble des paramètres temporels (c.f. section 2.3.3), acoustiques (section 2.3.4) et prosodiques (section 2.3.5). Un locuteur est alors représenté par un vecteur de paramètres calculés uniquement à partir des informations qui le concernent. C'est un point qui nous distingue des autres approches de l'état de l'art car cela permet de comparer des locuteurs issus de documents

audio différents. Ces descripteurs ne sont liés ni à la durée du document, ni à la position temporelle des segments des locuteurs dans le document. C'est un choix qui permet de représenter les locuteurs indépendamment de la structure du document. Nous ne voulons pas en effet exploiter la structure particulière d'un document pour trouver les rôles mais au contraire, pouvoir par la suite exploiter les rôles pour dégager les éléments caractéristiques de la structure des documents.

La suite de cette section s'organise de la manière suivante. Nous allons tout d'abord rappeler le format des résultats produits par un outil de SRL. Puis nous nous concentrerons sur la présentation des paramètres temporels, acoustiques et prosodiques sur lesquels s'appuie notre approche.

2.3.2 Segmentation et regroupement en locuteurs

La segmentation et regroupement en locuteurs dont le principe a été présenté dans la section 1.3.2.2 du chapitre précédent, est la première étape nécessaire à notre approche.

La SRL [El Khoury 09] permet de détecter les instants de parole d'un ensemble L de M locuteurs tel que $L = \{loc_1, loc_2, \dots loc_M\}$. Le nombre M de locuteurs n'est pas connu *a priori*.

À un locuteur donné, noté loc_j avec $j = 1 \dots M$, correspond un ensemble S_{loc_j} de N_j segments soit $S_{loc_j} = \{s_{(1,loc_j)}, s_{(2,loc_j)}, \dots, s_{(N_j,loc_j)}\}$. Chacun des segments est repéré par ses instants de début et de fin.

Sur la figure 2.2 est représenté un exemple de résultat d'une SRL. L'algorithme a détecté 6 locuteurs différents. Pour plus de lisibilité, sur la figure, une couleur correspond à un locuteur. Nous pouvons voir en particulier qu'une zone du signal audio n'a été attribuée à aucun locuteur. Le signal sur cette zone a été détecté comme ne correspondant pas à de la parole.

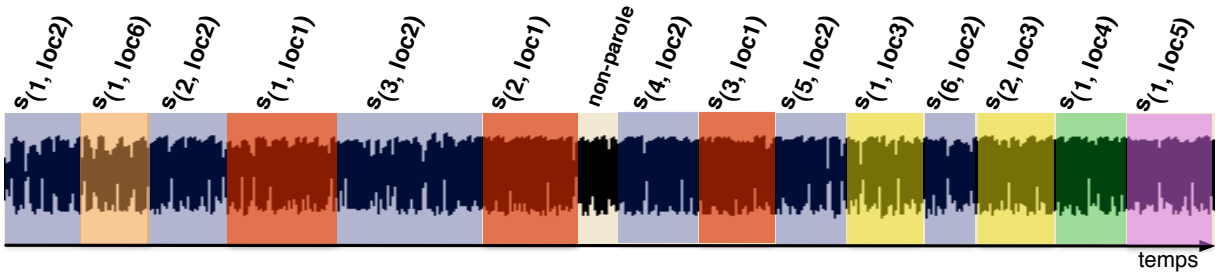


FIGURE 2.2 – Résultat d'une segmentation et regroupement en locuteurs

Les paramètres sont extraits des segments de chaque locuteur pris individuellement. Pour la suite, nous allégeons la notation en ne considérant plus les indices j et loc_j dans les équations. Pour un locuteur, nous disposons de N_{seg} segments $S = \{s_1, s_2, \dots, s_{N_{seg}}\}$. Pour chaque segment s_i, nous noterons D_i et F_i, ses instants de début et de fin.

2.3.3 Paramètres temporels

Les 14 paramètres temporels d'un intervenant sont calculés à partir des résultats de la SRL. Sur la figure 2.3 sont représentés uniquement les segments de parole du locuteur loc_2.

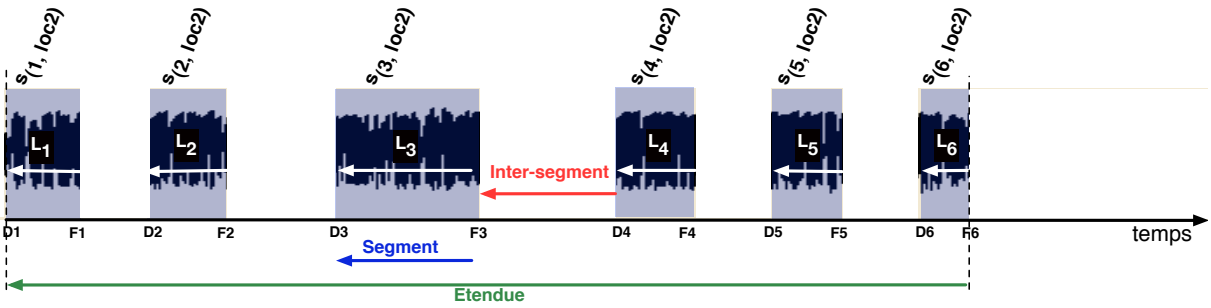


FIGURE 2.3 – Illustration avec les segments du locuteur loc_2 : longueur de segment, inter-segment et étendue.

2.3.3.1 Taille des segments et des inter-segments

Huit paramètres sont calculés à partir des longueurs des segments et des inter-segments du locuteur. L'inter-segment d'indice i correspond à l'intervalle entre la fin du segment d'indice i et le début du segment d'indice $i+1$ comme indiqué dans la figure 2.3. Les paramètres calculés à partir de ces deux mesures sont rassemblés dans la table 2.1. Il s'agit de la moyenne, de la variance, du maximum et du minimum.

TABLE 2.1 – Paramètres calculés à partir des segments et des inter-segments d'un locuteur comptant N_{seg} segments.

	À partir des longueurs de segments L_i	À partir des des inter-segments I_i
définition	$L_i = F_i - D_i$	$I_i = D_{i+1} - F_i$
moyenne	$\overline{L_{seg}} = \frac{1}{N_{seg}} \sum_{i=1}^{N_{seg}} L_i$	$\overline{I_{seg}} = \frac{1}{N_{seg}-1} \sum_{i=1}^{N_{seg}-1} I_i$
variance	$var(L_{seg}) = \frac{1}{N_{seg}} \sum_{i=1}^{N_{seg}} (L_i - \overline{L_{seg}})^2$	$var(I_{seg}) = \frac{1}{N_{seg}-1} \sum_{i=1}^{N_{seg}-1} (I_i - \overline{I_{seg}})^2$
maximum	$max(L_{seg}) = \max_{i=1}^{N_{seg}} L_i$	$max(I_{seg}) = \max_{i=1}^{N_{seg}-1} I_i$
minimum	$min(L_{seg}) = \min_{i=1}^{N_{seg}} L_i$	$min(I_{seg}) = \min_{i=1}^{N_{seg}-1} I_i$

À ce premier ensemble, nous ajoutons 6 paramètres temporels complémentaires pour caractériser globalement l'intervention du locuteur. Certains de ces paramètres comme l'activité globale, l'étendue et le nombre de segments ont été présentés dans le chapitre II. Nous les rappelons ici.

2.3.3.2 Quantification de l'Activité du locuteur

L'activité globale A correspondant au temps de parole cumulé du locuteur :

$$A = \sum_{i=1}^{N_{seg}} L_i$$

L'étendue E du locuteur visible sur la figure 2.3 est la durée qui sépare son premier instant et son dernier instant de parole :

$$E = F_{N_{seg}} - D_1$$

Nous ajoutons également 3 paramètres résultant de la combinaison des paramètres précédents.

Le taux d'extinction T_{ex} du locuteur, qui est la proportion de l'étendue du locuteur pendant laquelle celui-ci ne parle pas. C'est une mesure de la densité de son intervention :

$$T_{ex} = \frac{E - A}{E}$$

Le rapport entre le nombre de segments et l'activité globale du locuteur NsA, représente le degré de fragmentation de l'activité du locuteur :

$$NsA = \frac{N_{seg}}{A}$$

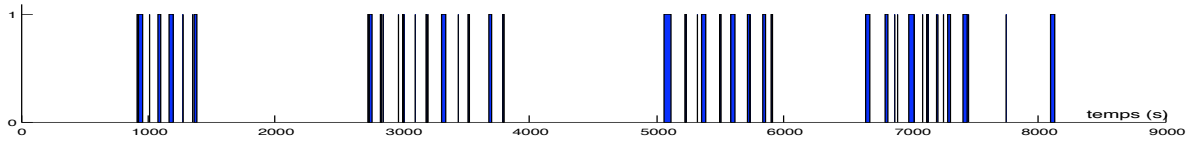
Le rapport entre le nombre de segments et l'étendue du locuteur NsE, similaire au paramètre précédent, il mesure le degré de fragmentation de la segmentation du locuteur par rapport à son étendue :

$$NsE = \frac{N_{seg}}{E}$$

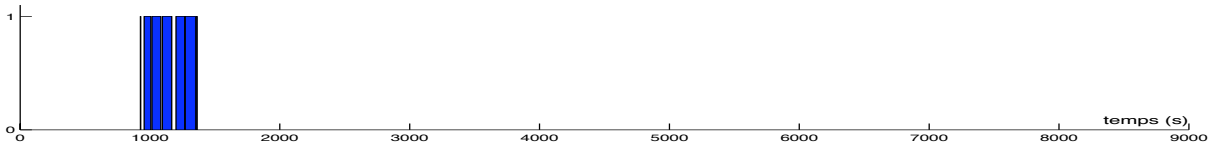
2.3.3.3 Les limites des paramètres temporels

La figure 2.4(a) représente les segments de parole d'un présentateur d'une émission d'information. La figure 2.4(b) représente les segments de parole d'un invité interviewé dans le cadre de cette émission. L'organisation temporelle des segments de parole de ces intervenants sont très différentes. Le présentateur est très actif, sa segmentation est plus étendue. L'intervention de l'invité est beaucoup plus dense et très localisée.

A contrario, à l'image des locuteurs loc_6 et loc_4 des figures 2.2 et 2.5, des locuteurs de rôles différents peuvent présenter des paramètres temporels similaires. Ces deux locuteurs ponctuels parlent pendant la même durée, ils présentent de fait des paramètres temporels rigoureusement identiques. Le locuteur loc_6 est un journaliste en duplex au téléphone, parlant dans un environnement très calme. L'audio correspondant est représenté la sous-figure 2.5(a). Le locuteur loc_4 est un anonyme interrogé en micro-trottoir en extérieur dans un environnement bruyant. L'audio correspondant à ce locuteur est représenté sur la sous-figure 2.5(b). Sur ces deux représentations,



(a) Les segments de parole d'un présentateur.



(b) Les segments de parole d'un invité.

FIGURE 2.4 – L'organisation temporelle des segments de parole (a) d'un présentateur et (b) d'un invité interviewé. Les instants de valeur égale à 1 indiquent quand le locuteur parle.

nous avons entouré d'ellipses rouges des zones durant lesquelles les locuteurs se taisent. Nous voyons que les signaux de chaque locuteur sont très différents l'un de l'autre. L'amplitude du signal sur ces zones sans parole pour le cas du locuteur loc_4, en milieu bruyant, est plus élevée que pour le locuteur loc_6 en milieu calme.

Nous souhaitons prendre en compte ce type d'informations acoustiques et observer si elles peuvent nous aider à discriminer les rôles des intervenants. Dans ce sens, nous proposons dans la suite de nouveaux paramètres extraits des informations acoustiques des interventions des locuteurs.

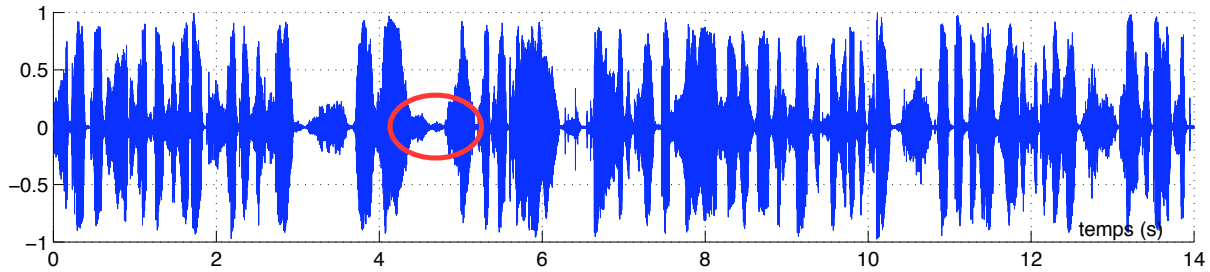
2.3.4 Paramètres acoustiques

Dix paramètres acoustiques sont calculés pour chaque locuteur à travers une analyse directe du signal audio correspondant. Certains locuteurs interviennent dans des environnements calmes quand d'autres intervenants sont enregistrés dans des environnements bruyants. Une manière basique de mesurer des caractéristiques de l'audio est d'observer les variations du signal dans le domaine temporel [Lu 98]. De cette manière, des statistiques sur l'intervention d'un locuteur peuvent être aisément obtenues. Nous proposons dans un premier temps de calculer des paramètres acoustiques sur la totalité de l'intervention du locuteur. Puis, à travers d'autres paramètres, nous évaluerons séparément les contributions énergétiques de l'activité parole du locuteur et de l'environnement séparément.

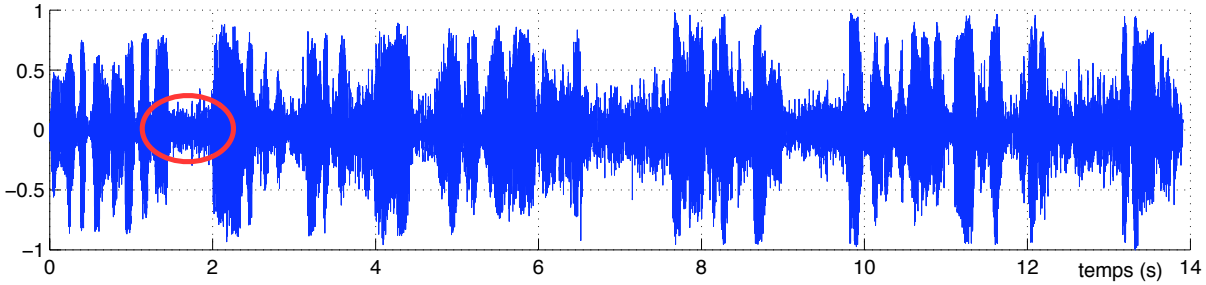
Calcul de la puissance moyenne du signal

Le signal de parole échantillonné $S_{loc_j}(k)$ avec $k = 1, \ldots K$ est obtenu à partir des segments audio du locuteur loc_j. Ici k est l'indice d'échantillon du signal S_{loc_j} de longueur K échantillons. Les paramètres acoustiques sont calculés pour chaque locuteur (dans la suite nous n'utiliserons plus l'indice loc_j du locuteur pour alléger la notation).

Le signal est centré et normalisé en amplitude. Compte tenu de la nature fortement non stationnaire du signal de parole, nous calculons la puissance court-terme du signal S sur des



(a) Le signal d'un journaliste loc_6.



(b) Le signal d'un interviewé en micro-trottoir loc_4.

FIGURE 2.5 – Les signaux audio (a) d'un journaliste, (b) d'un locuteur interviewé durant un micro-trottoir.

fenêtres glissantes de durée T échantillons tous les M échantillons. Nous réalisons de cette manière un lissage fin du signal.

Nous posons α l'indice de la fenêtre et n_α le premier échantillon de cette fenêtre. La puissance du signal $P(\alpha)$, sur la $\alpha^{ième}$ fenêtre est obtenue par la formule :

$$P(\alpha) = \frac{1}{T} \sum_{i=n_\alpha}^{n_\alpha+T-1} S(i)^2$$

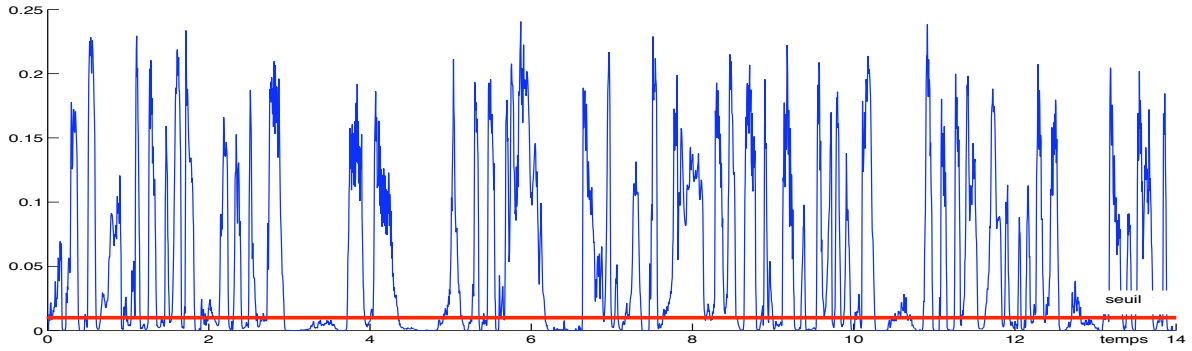
Nous choisissons des fenêtres de durée 10 millisecondes avec un recouvrement de 50%. Avec un signal audio échantillonné à une fréquence F_e, $T = 10.10^{-3} \times F_e$ et $M = 5.10^{-3} \times F_e$.

2.3.4.1 Puissance du signal sur l'intervention complète

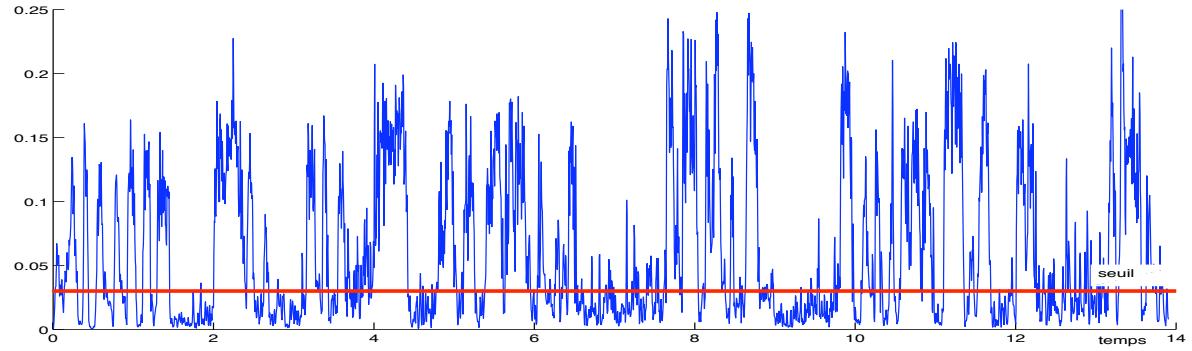
Nous calculons 2 paramètres à partir de la puissance du signal. Il s'agit de la valeur moyenne et de la variance de ce signal. Les calculs sont détaillés ci-dessous. Nous posons L la longueur du vecteur de la puissance du signal :

$$\text{moyenne de la puissance du signal} \quad \bar{P} = \frac{1}{L} \sum_{\alpha=1}^{L} P(\alpha)$$

$$\text{variance de la puissance du signal} \quad var(P) = \frac{1}{L} \sum_{\alpha=1}^{L} (P(\alpha) - \bar{P})^2$$



(a) L'amplitude de la puissance moyenne fenêtrée pour un journaliste parlant dans un environnement non buyant



(b) L'amplitude de la puissance moyenne fenêtrée d'un locuteur interviewé durant un micro-trottoir parlant dans un environnement bruyant

FIGURE 2.6 – Les courbes de la puissance du signal des interventions (a) d'un journaliste, (b) d'un locuteur interviewé durant un mico-trottoir. Le seuil indiqué par une ligne horizontale permet d'assurer la présence de la parole.

2.3.4.2 Contributions énergétiques du locuteur et de l'environnement sonore

Nous proposons d'extraire d'autres descripteurs pour mesurer plus spécifiquement les contributions énergétiques de la parole du locuteur et de l'environnement sonore. Dans l'absolu la séparation des contributions du locuteurs et de l'environnement sonore nécessite l'utilisation d'une approche fondée sur les méthodes de séparation aveugle de sources adaptées aux signaux de parole [Puigt 06].

Contrairement aux travaux de [Meinedo 03] nous ne souhaitons pas classer les environnements sonores dans des catégories précises. Nous souhaitons évaluer séparément les contributions énergétiques de la parole et de l'environnement sonore. Nous supposons que la voix du locuteur n'est pas complètement noyée dans le bruit, ainsi la contribution énergétique de la voix est plus importante que celle des bruits extérieurs (cas d'un rapport signal sur bruit positif). Sur la courbe $P(\alpha)$, nous utiliserons une approximation grossière en supposant que les zones d'amplitudes élevées correspondent aux instants où le locuteur parle.

Pour discriminer les zones avec et sans parole, nous nous inspirons des méthodes de détection d'activité vocale [Rabiner 75] basées sur l'analyse de l'énergie du signal. Par comparaison à un

seuil fixe égal à 15% de la puissance moyenne du signal ($\bar{P}$) nous positionnons les valeurs de $P(\alpha)$ dans ces catégories grossières. Ce seuil est matérialisé par une ligne horizontale sur les figures 2.6(a) et 2.6(b).

Nous calculons plusieurs paramètres pour les extraits du signal correspondant à ces deux catégories que nous assimilerons à la contribution du locuteur (marqués de l'indice *loc*) et à l'environnement sonore (avec l'indice *env*), soit un total de 8 paramètres :

- la valeur moyenne de la puissance du signal : $\overline{P_{loc}}$ et $\overline{P_{env}}$ dans chacun des cas,
- la variance de la puissance du signal : $var(P_{loc})$ et $var(P_{env})$,
- la valeur maximale et la valeur minimale du signal : $min(P_{loc})$, $min(P_{env})$, $max(P_{loc})$ et $max(P_{env})$.

2.3.5 Paramètres prosodiques

Pour compléter ce jeu de paramètres et tenir compte de l'élocution du locuteur, nous réalisons ensuite l'extraction de paramètres prosodiques.

Nous proposons 12 paramètres prosodiques que nous rassemblons en deux catégories. La première catégorie est liée à l'évolution de l'intonation et comprend plusieurs descripteurs calculés à partir d'une estimation de la fréquence fondamentale (F_0) du locuteur [de Cheveigné 02]. La seconde catégorie se concentre sur la mesure du débit de parole du locuteur et exploite pour cela les résultats d'une méthode de segmentation vocalique [Pellegrino 00].

2.3.5.1 Paramètres calculés à partir de la fréquence fondamentale

Quatre paramètres sont calculés à partir de l'estimation de la fréquence fondamentale. Nous utilisons l'algorithme Yin [de Cheveigné 02] pour ses performances et son temps de calcul réduit. Les courbes des figures 2.7(a) et 2.7(b) contiennent des représentations temporelles des estimations de la fréquence fondamentale, sur des extraits de durée 3 secondes, pour les mêmes locuteurs que précédemment. Nous pouvons y voir un certain nombre de zones non voisées, c'est à dire des zones sur lesquelles les cordes vocales du locuteur ne vibrent pas.

Les paramètres suivants sont calculés sur les zones où une fréquence fondamentale existe :

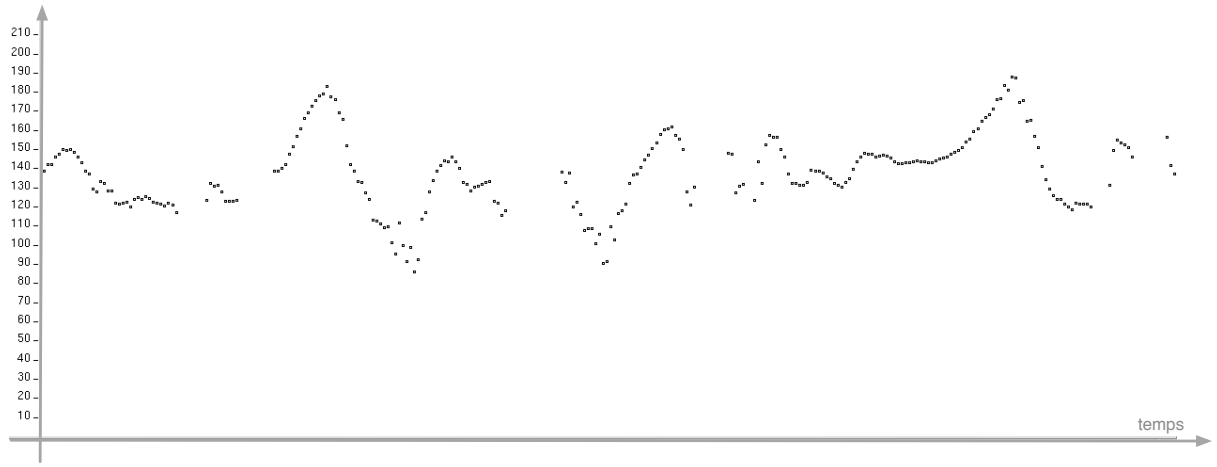
- la valeur moyenne de la fréquence fondamentale : $\overline{F_0}$,
- la variance de la fréquence fondamentale : $var(F_0)$,
- la valeur maximale de la fréquence fondamentale : $max(F_0)$.

Le quatrième paramètre correspond au **taux de zones voisées** T_{vois}, qui est le pourcentage du temps de parole total du locuteur durant laquelle la fréquence fondamentale existe.

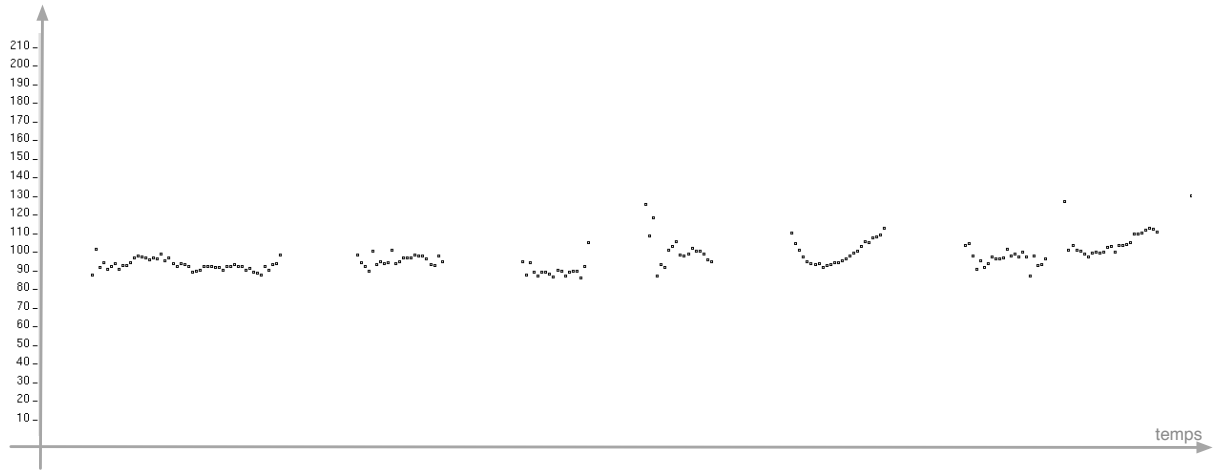
2.3.5.2 Paramètres calculés à partir d'une segmentation vocalique

La segmentation vocalique détecte les noyaux vocaliques et les silences d'un signal de parole. Nous utilisons une méthode [Pellegrino 00] développée dans notre équipe, basée sur une analyse spectrale du signal.

Nous voulons, à travers l'analyse du résultat de cette segmentation, évaluer la vitesse d'élocution d'un locuteur dont les variations peuvent être caractéristiques d'une intervention spontanée.



(a) L'évolution temporelle de fréquence fondamentale d'un journaliste



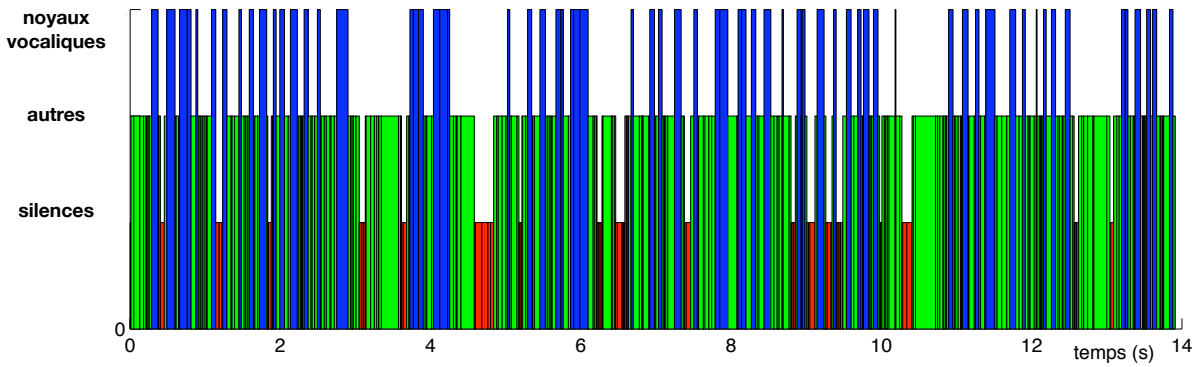
(b) L'évolution temporelle de la fréquence fondamentale d'un locuteur interviewé durant un micro-trottoir.

FIGURE 2.7 – Courbes représentant l'évolution de la fréquence fondamentale (a) d'un journaliste, (b) d'un locuteur interviewé durant un micro-trottoir.

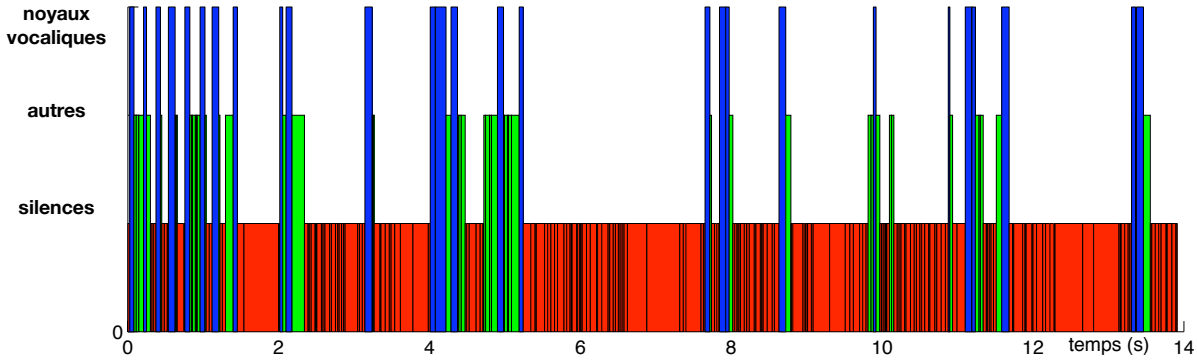
L'évaluation du nombre de noyaux vocaliques produits par unité de temps est une bonne approximation du débit de parole dans la mesure où elle représente une évaluation du nombre de syllabes par unité de temps.

Nous calculons 8 paramètres à partir du résultat de cette segmentation :

- le nombre de noyaux vocaliques N_{voy} et le nombre de silences N_{sil},
- le nombre de noyaux vocaliques par unité de temps $Debit_{voy}$ et le nombre de silences par unité de temps $Debit_{sil}$,
- la durée moyenne des noyaux vocaliques $\overline{Duree_{voy}}$ et la durée moyenne des silences $\overline{Duree_{sil}}$,
- la variance de la durée des noyaux vocaliques $var(Duree_{voy})$ et la variance sur la durée des silences $var(Duree_{sil})$.



(a) Positions des noyaux vocaliques, consonnes et silences issus d'une segmentation vocalique sur un extrait sonore d'un journaliste.



(b) Positions des noyaux vocaliques, consonnes et silences issus d'une segmentation vocalique appliquée sur un extrait sonore d'un locuteur interviewé durant un micro-trottoir.

FIGURE 2.8 – Résultats d'une segmentation vocalique appliquée à l'audio (a) d'un journaliste, (b) d'un locuteur interviewé durant un micro-trottoir. Trois types d'événements sont indiqués : en rouge les silences, en vert les zones complémentaires parmi lesquelles figures les zones consonnantiques et en bleu les noyaux vocaliques.

Les figures 2.8(a) et 2.8(b) représentent les résultats de la segmentation vocalique appliquée aux locuteurs pris comme exemples. Elles présentent de grandes différences concernant le nombre et la durée des voyelles, consonnes et silences. La segmentation du journaliste présente un grand nombre de voyelles et des silences courts. La segmentation du second locuteur présente au contraire de longs silences et peu de voyelles pour le même intervalle de temps.

2.3.6 Tableau de synthèse des paramètres

Nous avons rassemblé l'ensemble des paramètres temporels, acoustiques et prosodiques dans la table 2.2. Ces calculs sont réalisés pour chaque individu qui est alors représenté par un vecteur de 36 paramètres.

Nous allons dans la section suivante évaluer la pertinence des paramètres temporels, acoustiques et prosodiques en vue d'une utilisation dans un système de reconnaissance automatique des rôles.

TABLE 2.2 – L'ensemble des paramètres temporels, acoustiques et prosodiques et leurs symboles.

paramètres temporels	
longueur moyenne des segments	$\overline{L_{seg}}$
variance de la longueur des segments	$var(L_{seg})$
longueur max des segments	$max(L_{seg})$
longueur min des segments	$min(L_{seg})$
longueur moyenne des inter-segments	$\overline{I_{seg}}$
variance de la longueur des inter-segments	$var(I_{seg})$
longueur max des inter-segments	$max(I_{seg})$
longueur min des inter-segments	$min(I_{seg})$
Nombre de segments	N_{seg}
Activité globale	A
Etendue	E
Taux d'extinction	T_{ex}
nombre segment sur activité globale	NsA
nombre segments sur étendue	NsE
paramètres acoustiques	
puissance moyenne du signal	$\overline{P}$
variance de la puissance du signal	$var(P)$
puissance moyenne du signal sur les zones de parole	$\overline{P}_{loc}$
variance de puissance du signal sur les zones de parole	$var(P_{loc})$
valeur minimale de la puissance du signal sur les zones de parole	$min(P_{loc})$
valeur maximale de la puissance du signal sur les zones de parole	$max(P_{loc})$
puissance moyenne du signal sur les zones de non-parole	$\overline{P}_{env}$
variance de puissance du signal sur les zones de non-parole	$var(P_{env})$
valeur minimale de la puissance du signal sur les zones de non-parole	$min(P_{env})$
valeur maximale de la puissance du signal sur les zones de non-parole	$max(P_{env})$
paramètres prosodiques	
valeur moyenne de la fréquence fondamentale	$\overline{F_0}$
variance de la fréquence fondamentale	$var(F_0)$
valeur maximale de la fréquence fondamentale	$max(F_0)$
taux de zones voisées	T_{vois}
nombre de noyaux vocaliques	N_{voy}
nombre de noyaux vocaliques par seconde	$Debit_{voy}$
durée moyenne des noyaux vocaliques	$\overline{Duree_{voy}}$
variance sur la durée de noyaux vocaliques	$var(Duree_{voy})$
nombre de silences	N_{sil}
nombres de silences par unité de temps	$Debit_{sil}$
durée moyenne des silences	$\overline{Duree_{sil}}$
variance sur la durée de silences	$var(Duree_{sil})$

2.4 Validation des paramètres par une approche non supervisée

2.4.1 Objectif

Nous souhaitons évaluer la pertinence des paramètres temporels, acoustiques et prosodiques en vue d'une catégorisation des locuteurs parmi les 5 types de rôles : *présentateur, journaliste non ponctuel, autre non ponctuel, journaliste ponctuel et autre ponctuel*. Dans ce but, nous appliquons une méthode de regroupement non supervisé aux vecteurs de paramètres d'un ensemble de locuteurs. Nous espérons que les vecteurs de paramètres des locuteurs de même rôle, se rassemblent naturellement en clusters homogènes. Un tel résultat nous permettra de valider la pertinence des paramètres proposés et leur capacité à discriminer les différents rôles.

Le regroupement non supervisé est réalisé sur les vecteurs d'observation à l'aide de l'algorithme des K-means sur les corpus d'apprentissage et de test du corpus EPAC décrit dans la section 3.4.1.2.

Les vecteurs de paramètres sont extraits du corpus d'apprentissage à partir des informations disponibles après une segmentation manuelle en locuteurs. Les vecteurs de test sont issus d'un traitement entièrement automatique.

2.4.2 Méthode de regroupement non supervisé : algorithme des K-means

L'algorithme des K-means [Duda 00] est une approche de regroupement itérative qui rassemble les observations (dans notre cas des vecteurs de 36 paramètres) en K clusters, en fonction d'un critère de minimisation de distance au centre. Nous utilisons la distance euclidienne. Le nombre de clusters K est spécifié à l'initialisation de l'algorithme.

L'algorithme des K-means ne converge pas vers un optimum global et de ce fait le résultat du regroupement peut être différent d'une initialisation à une autre. L'utilisation de l'algorithme des K-means pour valider les paramètres se fait en deux étapes. Dans une première étape, nous utilisons le corpus d'apprentissage pour fixer le meilleur nombre K de clusters à partir d'un critère de stabilité de la répartition des observations en classes. Dans une seconde étape, l'algorithme des K-means est appliqué sur le corpus de test avec la valeur de K choisie et la qualité du regroupement est évaluée par rapport à la notion de rôle.

2.4.3 Critère de stabilité de la classification non supervisée

Nous présentons notre choix pour l'établissement de la meilleure valeur de K, le nombre initial de clusters du K-means. Soit $L_{app} = \{loc_1, \ldots, loc_M\}$, l'ensemble des M locuteurs du corpus d'apprentissage. Les rôles de ces intervenants ne sont pas connus. À l'aide de la segmentation manuelle en locuteurs, pour chaque intervenant loc_j de l'ensemble L_{app}, nous réalisons l'extraction des paramètres temporels, acoustiques et prosodiques. Un locuteur est ainsi représenté par un vecteur de 36 paramètres.

Notre critère de sélection de K, repose sur la recherche de la valeur de ce paramètres pour lequel le regroupement des locuteurs devient quasiment indépendant de l'initialisation. En d'autres termes, nous souhaitons trouver la valeur de K pour laquelle le résultat du regroupement est stable, quelque soient les observations utilisées pour initialiser l'algorithme. Dans la pratique

nous choisirons la première valeur de K pour laquelle 80% des initialisations génèrent le même partitionnement de l'espace des représentations des locuteurs.

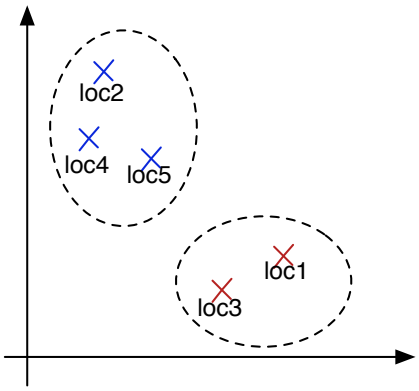
Nous testons des valeurs croissantes de K, en commençant avec $K = 5$, c'est à dire le nombre de types de rôles définies dans la section 2.2. Pour chaque valeur de K, nous réalisons 100 initialisations des K-means. Pour comparer les regroupements obtenus, nous étudions le contenu des clusters, et plus précisément nous observons « qui est regroupé avec qui ».

À l'issue de tout regroupement, nous remplissons une matrice d'appariement des locuteurs. Il s'agit d'une matrice de dimension $M \times M$. Les indices des lignes et des colonnes correspondent aux indices des identifiants des locuteurs loc_j avec $j = 1 \dots M$ de l'ensemble L_{app}. Nous remplissons la matrice de la manière suivante :

- si le locuteur loc_i et le locuteur loc_j appartiennent à un même cluster, la valeur correspondant à la $i^{ième}$ ligne et de $j^{ième}$ colonne vaut 1.
- cette valeur vaut 0 si les locuteurs appartiennent à des clusters différents.

La figure 2.9 donne une illustration d'une telle matrice.

Une matrice est générée après l'obtention du regroupement issu de chaque initialisation et au terme de la série d'initialisations, nous observons quel est le pourcentage d'initialisations ayant produit rigoureusement le même résultat. Avec des valeurs croissantes de K, nous observons la reproductibilité du regroupement, dans une proportion de 80% des cas, pour une valeur de **K égal à 20 clusters**.



	loc1	loc2	loc3	loc4	loc5
loc1	1	0	1	0	0
loc2	0	1	0	1	1
loc3	1	0	1	0	0
loc4	0	1	0	1	1
loc5	0	1	0	1	1

FIGURE 2.9 – Remplissage de la matrice d'appariement pour une initialisation avec K=2.

2.4.4 Analyse des résultats

Une manière de mesurer la qualité d'un regroupement, en terme de reconnaissance dans un problème multi-classes défini *a priori*, est de calculer la pureté des clusters [Duda 00], par rapport à ces classes.

La pureté du cluster C_j se mesure en fonction de la classe qu'il contient majoritairement.

$$pureté(C_j) = \frac{1}{|C_j|} \max_i(|C_j|_{classe=i})$$

$|C_j|_{classe=i}$ est le nombre d'éléments du cluster C_j appartenant à la classe i définie *a priori*.

La pureté moyenne sur l'ensemble des K clusters se calcule avec la formule suivante, où M est le nombre total d'observations.

$$\overline{pureté} = \sum_{j=1}^{K} \frac{|C_j|}{M} pureté(C_j)$$

Un cluster dont la pureté est supérieure à 50% est attribué à la classe majoritairement représentée. Nous calculons également **la pureté d'un rôle** en adaptant le calcul de la pureté moyenne aux clusters dans lesquels ce rôle est contenu majoritairement.

TABLE 2.3 – Étiquetage des clusters en rôles.

	nombre de clusters	pureté en rôle
présentateur	2	71%
journaliste non ponctuel	3	66%
autre non ponctuel	10	80%
journaliste ponctuel	4	87%
autre ponctuel	1	60%
$\overline{pureté}$ (écart-type)	80% (19%)	

Les performances du regroupement non supervisé pour une initialisation à $K = 20$ clusters sur le corpus de test EPAC sont rapportées dans la table 2.3. Dans cette table, nous indiquons le nombre de clusters attribués à chaque classe, la pureté moyenne de chaque rôle relative aux clusters associés et la pureté moyenne sur l'ensemble des clusters.

Globalement sur les vingt clusters,

- deux clusters contiennent en majorité des *présentateur* avec une proportion moyenne de 71%.
- La classe *journaliste non ponctuel* est majoritaire en moyenne à 66% dans 3 clusters.
- La classe *autre ponctuel* compte 10 clusters purs en moyenne à 80%.
- Les *journaliste ponctuel* sont majoritaires dans 4 clusters à 87% de pureté en moyenne.
- Enfin, la classe *autre ponctuel* est attribuée à un seul cluster pur à 60%.

Aucune classe n'est perdue lors du regroupement, puisque à un rôle correspond au moins un cluster.

Nous observons également que le nombre de clusters n'est pas équitablement réparti parmi les 5 classes. Ces nombres doivent être confrontés au contenu réel, c'est à dire à la variété de locuteurs qui peuvent être rassemblés sous une même étiquette de rôle.

Nous avions identifié, à travers la description des rôles présentées dans la section 2.2 que les classes *autre non ponctuel* et *autre ponctuel* sont celles qui rassemblent sous une même étiquette des intervenants présentant des caractéristiques *a priori* les plus variées.

Cette observation semble être confirmée par les résultats de regroupement non supervisé. La classe *autre non ponctuel* est reconnue avec une pureté élevée, mais les locuteurs de cette classes sont dispersés parmi 10 clusters. Ce nombre de clusters est élevé par rapport aux autres rôles. L'extrême inverse est observé pour la classe *autre ponctuel* associée à un seul cluster, dont la pureté est plutôt faible.

Les classes *présentateur*, *journaliste non ponctuel* et *journaliste ponctuel* sont majoritaires respectivement dans 2, 3 et 4 clusters. Ces nombres sont *a priori* cohérents avec l'idée que nous nous faisons de la variété de locuteurs rassemblés dans ces catégories de rôles.

Finalement ce résultat très encourageant nous permet d'envisager l'intégration de ces paramètres dans un système de reconnaissance automatique des rôles présenté dans le chapitre suivant.

2.5 Conclusion

Dans ce chapitre nous avons présenté une étude ayant pour objectif la recherche de paramètres pertinents en perspective du développement d'un système de reconnaissance automatique de rôles. La littérature présente plusieurs approches de reconnaissance des rôles s'appuyant sur le contenu du message prononcé par les individus. Notre approche se fonde sur l'hypothèse qu'il existe des informations non lexicales sur les rôles locuteurs, disponibles dans un ensemble de paramètres bas-niveau. Dans un premier temps nous avons précisé les définitions des catégories de rôles recherchés : *présentateur*, *journaliste* et *autre*. Cet ensemble de trois rôles communs aux travaux de la littérature ont été étendus à un ensemble de 5 rôles grâce à une distinction faite entre les intervenants ponctuels et non ponctuels.

Nous avons ensuite proposé et décrit un ensemble de 36 paramètres « bas-niveau » se composant de 14 paramètres temporels permettant de caractériser l'organisation temporelle des tours de parole d'un locuteur, de 10 paramètres acoustiques dont le but est d'évaluer l'adaptation du locuteur à son environnement sonore, et d'un ensemble de 12 paramètres prosodiques concernant plus particulièrement l'étude de l'intonation et du débit de parole de l'intervenant.

La pertinence de paramètres a été évaluée, sur l'ensemble de documents audio du corpus EPAC, par le biais de l'analyse du résultat d'un regroupement non supervisé. Les clusters finalement obtenus rassemblent des locuteurs de rôles similaires et présentent une pureté moyenne de 80%.

Ce bon résultat nous permet d'envisager l'exploitation de cette représentation des rôles des locuteurs dans un système automatique de reconnaissance de rôles, présenté dans le chapitre suivant.

Chapitre 3

Système de reconnaissance automatique des rôles

Nous avons proposé dans le chapitre précédent trois ensembles de paramètres dits de « bas-niveau » (temporels, acoustiques et prosodiques), par le biais desquels nous espérons capturer la part d'information non lexicale caractérisant l'intervention de chaque locuteur. Les résultats de nos investigations précédentes obtenus à l'aide d'une classification non supervisée démontrent qu'il existe effectivement un lien entre l'information contenue dans ces paramètres et les rôles joués par les intervenants. Afin de confirmer ces hypothèses et d'en évaluer le potentiel, nous développons un système de recherche automatique des intervenants et de leur rôle dans un flux audio. Ce chapitre est consacré à sa présentation, avec une argumentation des choix effectués, étayée par une évaluation systématique.

3.1 Architecture d'un système de reconnaissance automatique des rôles des intervenants dans un flux audio

Ce système de reconnaissance automatique suit la structure classique d'un système de reconnaissance des formes [Duda 00] auquel il faut adjoindre une phase de segmentation préalable à même de produire les zones à partir desquelles la détection de rôle sera effectuée. L'architecture de ce système se décompose naturellement en quatre parties correspondant aux étapes de traitement suivantes (figure 3.1) :

- le flux audio est, tout d'abord, traité par un algorithme de segmentation et de regroupement en locuteurs (SRL), présenté dans la section 1.3.2.2,
- pour chaque locuteur détecté, l'extraction des paramètres temporels, acoustiques et prosodiques est réalisée. L'intervention de chaque locuteur est ainsi représentée par un ensemble de 36 paramètres. Cette étape est détaillée dans la section 2.3 et un tableau de synthèse est présenté à la fin du chapitre 2.
- une transformation éventuelle du vecteur de paramètres, résultant d'une étude portant sur la réduction de dimension, est faite en phase d'apprentissage,

– enfin, une étape de décision permet, à partir des paramètres pertinents retenus, d'associer un rôle à chaque locuteur détecté.

Au cours de ce chapitre sont présentées les méthodes que nous avons étudiées pour la réduction de dimension (section 3.2) et pour la prise de décision (section 3.3). Le protocole expérimental, où sont précisés les corpus et les mesures d'évaluation, fait l'objet de la section 3.4. Différentes configurations du système sont évaluées et leurs performances sont rassemblées dans les sections 3.5, 3.6 et 3.7. Des compléments à ces sections sont donnés dans l'annexe A.

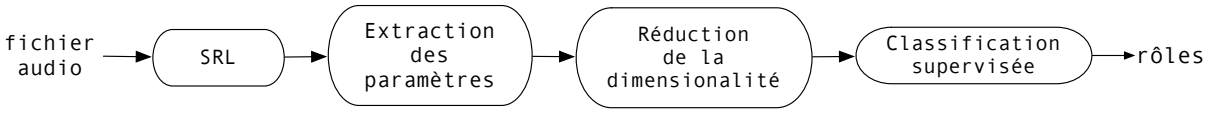


FIGURE 3.1 – Architecture du système de reconnaissance des rôles.

3.2 Méthodes de réduction de dimension

La réduction de dimension a été largement étudiée et un panorama des approches existantes peut être trouvé dans [Carreira-Perpiñán 97] et [Cunningham 08]. Ces méthodes sont généralement utilisées pour :

– débruiter les données, c'est-à-dire éliminer les données non significatives qui pourraient avoir un impact négatif sur les performances des classifieurs,
– affranchir les phases d'apprentissage des problèmes rencontrés lorsque le nombre de paramètres devient trop important par rapport au nombre d'observations (connu sous le nom de « fléau de la dimension » [Bellman 61]),
– réaliser une meilleure analyse des données, en ne conservant que les paramètres les plus discriminants,
– réduire, sans perte d'information, le coût calculatoire du traitement des données.

Les 36 paramètres, qui caractérisent chaque locuteur détecté, sont d'une part, très corrélés et d'autre part, ont été définis intuitivement. Afin de limiter l'influence sur la reconnaissance des rôles des paramètres corrélés ou porteurs de peu d'information, nous étudions donc la possibilité d'appliquer une réduction de dimension aux vecteurs de données. Cette étude s'impose d'autant plus que le nombre de paramètres est effectivement assez élevé et, comme cela sera précisé dans la section 3.4, le volume du corpus d'apprentissage reste modeste.

Les méthodes de réduction de dimension peuvent être de deux types : les approches par transformation de paramètres (présentées dans la section 3.2.1) et les approches par sélection de paramètres (section 3.2.2). Chaque type d'approche peut être envisagé en mode supervisé ou non supervisé, selon que l'on dispose ou non d'une information sur la classe des données d'apprentissage. Ceci est illustré par la figure 3.2. Au cours des deux paragraphes suivants, nous limitons la présentation aux seules méthodes utilisées dans notre système.

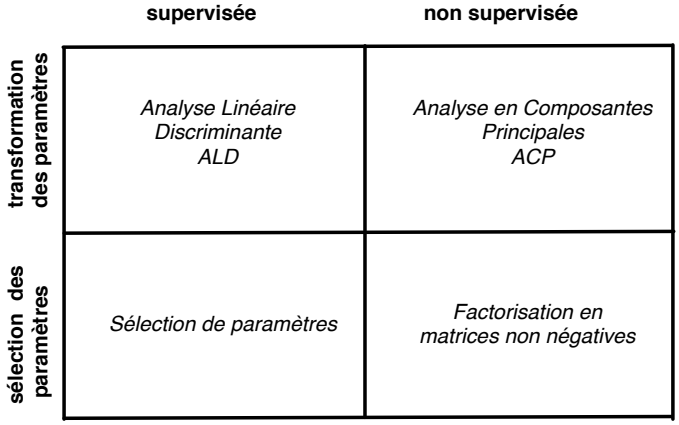


FIGURE 3.2 – Tableau récapitulatif des principales méthodes de réduction de dimension en fonction du type de transformation et du mode d'apprentissage.

Afin de faciliter la compréhension, nous utilisons les notations suivantes. L'ensemble des données d'apprentissage $\mathbf{A}$ est composé de n observations, chacune est de dimension p. Une méthode de réduction de dimension a pour objectif de fournir une application qui transforme l'ensemble initial des données en un nouvel ensemble $\mathbf{A}'$ composé de n observations, chacune étant de dimension k avec $(k < p)$, en conservant la plus grande part de l'information initiale contenue dans $\mathbf{A}$, sous contrôle d'un critère.

3.2.1 Réduction de dimension par transformation des paramètres

Dans cette catégorie de méthodes, les données de l'ensemble $\mathbf{A}$ sont projetées dans un nouvel espace de représentation (de dimension inférieure) par l'application d'une transformation linéaire telle que :

$$\mathbf{A}' = \mathbf{MA}$$

La matrice de transformation $\mathbf{M}$ peut être calculée par une méthode non supervisée comme l'Analyse en Composantes Principales (ACP) ou supervisée comme l'Analyse Factorielle Discriminante (AFD). Nous détaillons ces deux méthodes que nous utiliserons lors des expérimentations.

L'Analyse en Composantes Principales (ACP) est une méthode d'analyse de données qui réalise une projection d'un ensemble de données de dimension élevée dans un nouvel espace de dimension plus réduite dont les axes sont appelés les axes principaux [Duda 00]. Les coordonnées d'une donnée projetée sont des combinaisons linéaires des coordonnées initiales et sont appelées les composantes principales de cette donnée. La transformation linéaire est obtenue par maximisation de l'inertie du nuage des données projetées. Le nombre de composantes principales retenues est calculé de manière à représenter une proportion maximale de la variance des données. Le processus est itératif : le premier axe correspond à la direction qui maximise la variance des données projetées, le second est un axe orthogonal au premier qui maximise la variance restante et ainsi de suite. La matrice de transformation linéaire est la matrice de passage entre

la base canonique et les vecteurs propres de la matrice de covariance de l'ensemble de données initial. La réduction de dimension est réalisée en conservant seulement les k axes principaux les plus importants, représentant une large proportion de la variance des données. Dans la pratique, nous conservons les axes principaux contenant 99% de cette variance.

L'Analyse Factorielle Discriminante (AFD) est une méthode de réduction de dimension qui préserve les classes auxquelles appartiennent les données [Fisher 36]. C'est pourquoi cette méthode s'effectue en mode supervisé car l'appartenance des données à un ensemble de classes est connue. La matrice $\mathbf{M}$ projette les données de $\mathbf{A}$ dans un espace tout en maximisant l'inertie inter-classe et minimisant l'inertie intra-classe. Comme pour l'ACP, il s'agit d'une projection et les coordonnées dans le nouvel espace de représentation sont des combinaisons linéaires des coordonnées initiales. Les propriétés mathématiques des matrices d'inertie intra-classe et inter-classe obligent à conserver au plus $(C-1)$ axes discriminants, où C correspond au nombre de classes initiales.

3.2.2 Réduction de dimension par sélection de paramètres

Ces méthodes visent à isoler le sous-ensemble de paramètres le plus discriminant dans une tâche particulière de classification. Les données ne sont pas transformées, au contraire des méthodes précédentes, elles sont conservées dans leur espace initial, permettant ainsi une interprétation directe des résultats. Nous n'envisagerons que les méthodes supervisées pour lesquelles la classe de chacune des données d'apprentissage est connue, comme cela sera notre cas.

Plusieurs stratégies sont alors possibles :

- certaines sont basées sur un calcul de critères objectifs (test du χ^2, gain en information), évalués sur l'ensemble de données $\mathbf{A}$ [Wu 02]. Les paramètres les moins discriminants au sens de ces critères sont écartés.
- d'autres font appel aux méthodes de classification par arbres de décision [Quinlan 93]. Chaque noeud de l'arbre conduit à une règle et à la sélection d'un paramètre. Ces méthodes mènent à un ensemble de règles qui peuvent être utilisées pour classer de nouvelles données.
- enfin, une troisième catégorie de méthodes sélectionne les paramètres pertinents en couplant sélection et classification. Nous détaillons plus particulièrement cette catégorie dans la suite.

La sélection de paramètres par validation en classification vise à rechercher le jeu de paramètres permettant d'atteindre la meilleure performance en terme de taux de reconnaissance pour une méthode de classification donnée [Cunningham 08]. Chaque sous-ensemble de paramètres est évalué par validation croisée sur l'ensemble des données d'apprentissage $\mathbf{A}$. En faisant varier également la méthode de classification, cette approche permet de sélectionner le couple (jeu de paramètres/méthode de classification) qui obtient les meilleures performances, couple qui sera par la suite utilisé pour classer de nouvelles données.

Cette approche implique une combinatoire de jeu-test élevée. Le seul nombre de combinaisons de paramètres à examiner pour une méthode de classification croît en 2^p, avec p le nombre de paramètres. Pour cette raison, une recherche exhaustive n'est souvent pas envisageable. Plusieurs stratégies sont possibles et nous avons retenu la méthode de sélection par élimination, nommée

Recherche Séquentielle par Élimination (RSE)[Guyon 03]. Il s'agit d'une procédure de recherche par retrait successif d'un paramètre à partir de l'ensemble complet des paramètres. La variable ôtée est celle dont le retrait dégrade le moins les performances de classification. Le procédé est arrêté soit quand il ne reste qu'un seul paramètre dans l'ensemble des variables de départ, soit lorsque que le taux de reconnaissance est trop fortement dégradé.

3.3 Méthodes de classification supervisée

De manière similaire aux méthodes de réduction de dimension, il existe deux types d'approche : l'approche supervisée et l'approche non-supervisée. Nous consacrons la suite de cette section à la présentation des approches de classification en mode supervisé, puisque tel est le cadre de notre étude. Le lecteur pourra se référer à [Duda 00] pour de plus amples informations sur les approches non-supervisées.

Rappelons que l'approche de classification est dite « supervisée » lorsque l'appartenance de chaque donnée d'apprentissage à un élément de l'ensemble C, composé de I classes, est connue et est utilisée pour établir les fonctions de décision (modèles probabilistes ou règles) permettant de prédire l'appartenance de chaque nouvelle observation à l'une de ces classes. Nous détaillons maintenant les trois types de méthodes de classification supervisée que nous utilisons dans la suite de ce chapitre.

3.3.1 Modèles de Mélanges de lois Gaussiennes (GMM)

Nous noterons dans la suite A_{C_i} le sous-ensemble des données de l'ensemble d'apprentissage A appartenant à la classe C_i et y, x des données d'observation. Dans le cadre d'une modélisation par mélange de lois Gaussiennes, les observations appartenant à chaque classe C_i sont supposées suivre une loi de type GMM, un GMM étant une somme pondérée de lois Gaussiennes ou normales comme schématisé dans la figure 3.3.

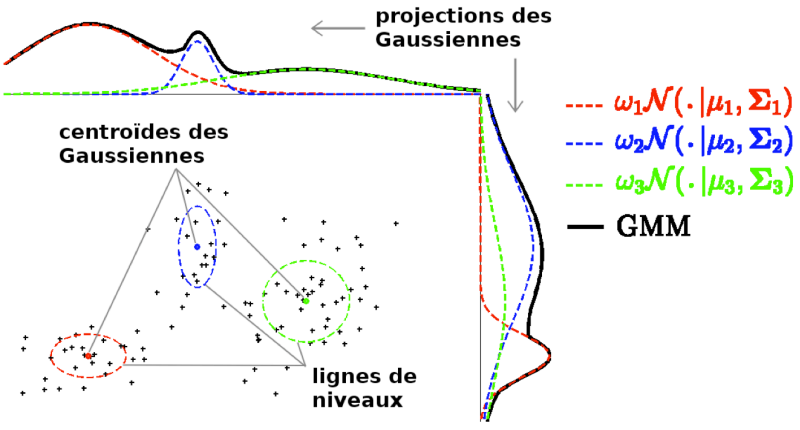


FIGURE 3.3 – Visualisation d'un ensemble d'observations distribuées selon un mélange de trois lois Gaussiennes (exemple en deux dimensions).

La densité de probabilité des observations y, appartenant à la classe C_i, pour un mélange de N Gaussiennes, peut être écrite sous la forme suivante :

$$p(y|\Theta_i) = \sum_{n=1}^{N} \alpha_n^i \mathcal{N}(y, \mu_n^i, \Sigma_n^i)$$

$$\text{avec } \Theta_i = \{\alpha_1^i, \mu_1^i, \Sigma_1^i, \ldots, \alpha_N^i, \mu_N^i, \Sigma_N^i\} \text{ et } \sum_{n=1}^{N} \alpha_n^i = 1$$

En phase d'apprentissage, il est nécessaire, pour chaque composante gaussienne n du mélange, d'estimer l'ensemble des paramètres μ_n^i, Σ_n^i et α_n^i qui sont respectivement la moyenne, la variance et le coefficient de pondération de cette $n^{ième}$ composante gaussienne. Ces estimations sont réalisées par l'algorithme d'Espérance-Maximisation [Dempster 77].

En phase de reconnaissance, l'attribution d'une nouvelle donnée x à l'une des classes de l'ensemble C est accomplie en choisissant la classe C_i, parmi I classes possibles, qui maximise le maximum de vraisemblance. Λ_x est appelé l'estimateur par maximum de vraisemblance :

$$\Lambda_x = \arg \max_{i=1\ldots I} p(x|\Theta_i)$$

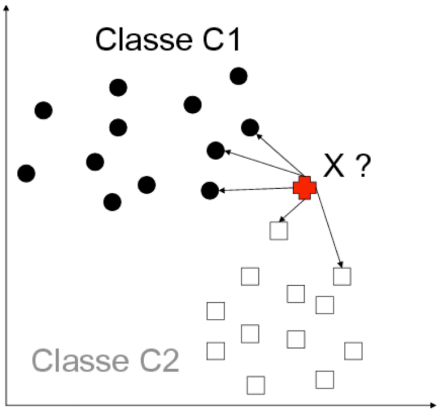


FIGURE 3.4 – Illustration de la méthode des k-plus proches voisins : cas d'un problème à deux classes C1 et C2 avec k=5 et X la donnée à classer (avec ici X plus proche de C1).

3.3.2 Méthode des k-plus proches voisins (k-ppv)

La méthode des k-ppv repose sur l'estimation locale des densités de probabilité [Duda 00]. Aucun modèle n'est calculé en phase d'apprentissage et, en phase de reconnaissance, les distances entre l'observation à classer et les observations rassemblées lors de l'apprentissage **A** sont mesurées. La classe la plus représentée, parmi les k observations d'apprentissage les plus proches de l'échantillon à classer, est attribuée à ce dernier (cf. exemple de la figure 3.4).

Outre sa simplicité, l'avantage de cette méthode est qu'elle peut naturellement s'appliquer au cas multi-classes, même avec un nombre élevé de classes. Mais elle présente aussi un désavantage

dans la mesure où un volume important de données d'apprentissage implique d'une part la nécessité de disposer d'une capacité mémoire d'autant plus élevée, et entraîne d'autre part une forte complexité calculatoire en phase de test.

3.3.3 Machines à Vecteurs de Support (SVM)

Les méthodes de type SVM sont des méthodes de classification formulées pour résoudre des problèmes à deux classes [Vapnik 98]. Le but des SVM est alors de trouver un classifieur qui sépare les données en maximisant la distance entre ces deux classes. Dans le domaine linéaire, cela revient à trouver un hyperplan séparateur optimal. Pour chaque problème de classification linéaire, il existe une multitude d'hyperplans valides, mais dans la démarche SVM, il s'agit de rechercher un hyperplan dont la distance minimale aux exemples d'apprentissage se trouve être maximale. Cette distance s'apelle la « marge » et l'hyperplan séparateur optimal est celui qui maximise la marge. Deux exemples d'hyperplans possibles sont présentés sur la figure 3.5.

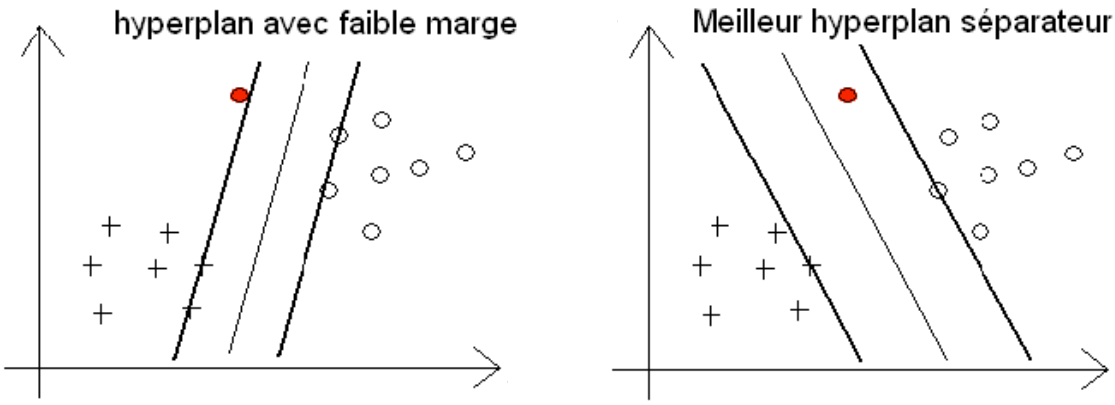


FIGURE 3.5 – Deux exemples d'hyperplan séparateur avec des marges différentes.

Il est assez rare que les données soient directement séparables par un hyperplan dans l'espace de représentation initial. Pour traiter les cas non linéairement séparables, les SVM exploitent « l'astuce du noyau » qui conduit à plonger simplement les données dans un espace de dimension supérieure où elles sont alors séparables par un hyperplan. Il existe un certain nombre de noyaux connus s'ajoutant à la forme originelle linéaire. Les noyaux d'usage courant sont rappelés dans [Ramona 10]. Nous utilisons dans ce manuscrit les noyaux suivants, où x et y sont des vecteurs d'observations, c, d et σ, les paramètres des méthodes :

- le noyau linéaire : $k(\mathbf{x},\mathbf{y}) = \mathbf{x}^T\mathbf{y}$ qui correspond au produit scalaire dans l'espace initial sans opérer de transformation. C'est la forme la plus traditionnelle.
- le noyau polynomial non homogène : $k(\mathbf{x},\mathbf{y}) = \big(1 + \frac{c}{d}\mathbf{x}^T\mathbf{y}\big)$,
- le noyau gaussien RBF (Radial Basis Functions) : $k(\mathbf{x},\mathbf{y}) = \exp\big(-\frac{||\mathbf{x}-\mathbf{y}||^2}{d\sigma^2}\big)$,
- le noyau sigmoïdal : $k(\mathbf{x},\mathbf{y}) = \tanh\big(\frac{c}{d}\mathbf{x}^T\mathbf{y} + \Theta\big)$.

3.4 Protocole expérimental

Afin de construire un système performant de reconnaissance automatique des rôles des intervenants, nous avons réalisé un grand nombre d'expérimentations visant à préciser la meilleure configuration en termes de paramètres, de stratégie de décision et de robustesse vis-à-vis des données elles-mêmes. Avant de présenter ces expériences accompagnées de leurs résultats, il convient de préciser leur cadre, à savoir la nature des corpus utilisés et les mesures d'évaluation.

3.4.1 Corpus

Comme dit en introduction générale, ce travail a été réalisé dans le cadre du projet EPAC. Les premières évaluations de notre système de reconnaissance des rôles sont antérieures à la livraison du corpus EPAC par les annotateurs. Nous avons donc réalisé nos premières expériences avec un autre corpus. Notre choix s'est porté sur le corpus ESTER2 car d'une part, son contenu est relativement proche de celui du corpus EPAC et d'autre part, l'annotation manuelle en locuteurs, coûteuse à produire, est déjà disponible sur ce corpus fourni dans le cadre de la campagne d'évaluation ESTER2, campagne au cours de laquelle le système de segmentation et regroupement en locuteurs que nous utilisons a été évalué. C'est pour ces différentes raisons, qu'une grande part des expériences réalisées porte sur le corpus ESTER2. Des expériences sur le corpus EPAC viendront compléter notre travail en fin de chapitre.

3.4.1.1 Corpus ESTER2

Ce corpus est un ensemble de documents audio correspondant à des enregistrements radiophoniques produits pour la campagne d'évaluation ESTER2 [Galliano 09]. Pour l'évaluation des différents systèmes de reconnaissance de rôles que nous proposons, nous utilisons les ensembles de développement et de test, nommés respectivement ESTER-dev et ESTER-tst. Le premier ensemble représente 6 heures de données audio et est utilisé ici comme ensemble d'apprentissage, tandis que le second qui représente 7 heures de données audio est utilisé comme ensemble de test. À eux deux, ESTER-dev et ESTER-tst rassemblent 46 documents soit 13 heures d'enregistrements annotées manuellement en locuteurs.

Ces documents correspondent à 41 bulletins d'informations et 5 talk-shows, issus de plusieurs stations de radio francophones : TVME, Africa1, France Inter et RFI. Ce corpus contient une grande diversité en terme de programmes, de durées d'émission, de créneaux horaires, et de nombres de locuteurs par document. L'algorithme de segmentation et de regroupement en locuteurs que nous utilisons [El Khoury 09] rapporte une valeur de DER (Diarization Error Rate) d'environ 9% sur l'ensemble ESTER-tst. Pour les besoins de l'évaluation de nos différents systèmes, il a été nécessaire de compléter ces annotations manuellement pour y intégrer le rôle de l'intervenant. Il est rappelé que les classes considérées sont : *présentateur*, *journaliste* et *autre*, avec distiction ou non des locuteurs ponctuels.

Dans ce corpus, chaque document correspond à une seule émission et ne compte qu'un locuteur dont le rôle est *présentateur*. Dans la mesure où chaque vecteur d'observation correspond à un locuteur, il y a autant de vecteurs d'observation que de locuteurs trouvés manuellement. Un même locuteur, appartenant à deux documents différents, est considéré comme deux observations distinctes, qu'il soit investi du même rôle ou non, ce qui explique qu'il y ait 46 présentateurs

au total dans ce corpus. La répartition, en nombre de locuteurs ou d'observations pour chaque rôle considéré, est donnée dans la table 3.1 pour les ensembles ESTER-dev et ESTER-tst. Cette table montre que les différentes classes de rôle y sont représentées en proportions similaires.

TABLE 3.1 – Nombre et proportion de chaque rôle dans les ensembles ESTER-dev et ESTER-tst.

	présentateur	journaliste non ponct	autre non ponct	journaliste ponctuel	autre ponctuel	total
ESTER-dev	20(8%)	72(29%)	65(26%)	39(15%)	55(22%)	251(100%)
ESTER-tst	26(13%)	55(27%)	59(29%)	35(17%)	28(14%)	203(100%)

3.4.1.2 Corpus EPAC

Le corpus EPAC est extrait de l'ensemble plus large des 2000 heures enregistrées dans le cadre de la campagne ESTER mais majoritairement non annotées. Ces 2000 heures correspondent à plusieurs stations de radio généralistes francophones (France Inter, France Info, France Culture et RFI) et contiennent plusieurs catégories d'émissions comme des débats, des émissions matinales, des bulletins d'information ou des magazines. L'un des objectifs du projet EPAC, décrit dans la section 3 de l'introduction générale, a été d'annoter manuellement 100 heures issues de ce corpus en mettant l'accent sur la parole conversationnelle.

Les annotations manuelles ont été finalisées en 2010. Elles ont été réalisées à plusieurs niveaux de détails et contiennent entre autre :

- la segmentation en locuteurs,
- la transcription de la parole,
- le rôles des locuteurs,
- le thème des émissions,
- l'identité des locuteurs.

Afin de pouvoir réaliser des évaluations internes pour les différents systèmes développés par les partenaires du projet, les 100 heures de documents audio ont été séparées en trois sous-ensembles qui sont :

- EPAC-app : ensemble d'apprentissage de 80 heures,
- EPAC-dev : ensemble de développement de 10 heures,
- EPAC-tst : ensemble de test de 10 heures.

Dans nos propres expériences, nous avons réalisé l'étape d'apprentissage de nos classifieurs sur l'ensemble EPAC-app et la reconnaissance des rôles sur l'ensemble EPAC-tst. La SRL appliquée à l'ensemble de test EPAC-tst rapporte un DER égal à 7%.

À la différence du corpus ESTER, les documents du corpus EPAC peuvent contenir plusieurs émissions successives et donc plusieurs présentateurs différents. Il y a même des émissions animées par deux présentateurs. Par contre, il n'y a pas d'émission sans présentateur.

Le choix des documents annotés dans le cadre du corpus EPAC a été guidé par la présence de parole conversationnelle. Aussi, le corpus EPAC contient-il en majorité des émissions de société,

tels que des débats ou des magazines. De ce fait également, les bulletins d'information y sont beaucoup moins représentés que dans le corpus ESTER2. Le volume du corpus EPAC étant plus important, cela implique une plus grande variété des rôles des locuteurs. Les proportions de locuteurs appartenant à chaque classe sont également différentes. La table 3.2 rassemble le nombre de locuteurs et les proportions de chaque type de rôle présents dans ce corpus. Son examen nous amène à faire quelques remarques :

- Le rôle *présentateur* regroupe à la fois des présentateurs de journaux d'information, de matinales, d'émissions de société, de magazines culturels et de débats.
- Le rôle *journaliste* correspond plutôt à des chroniqueurs, des reporters, des correspondants à l'étranger, des analystes et des speaker(-ines) qui apparaissent plus souvent de manière ponctuelle. Le nombre de *journaliste ponctuel* est donc plus important que le nombre de *journaliste non ponctuel*.
- Le rôle *autre* recouvre un panel plus vaste de type d'intervenants qui sont plus ou moins habitués à parler à la radio, plus ou moins impliqués dans leurs interventions, plus ou moins hésitants. Les émissions de société font plutôt intervenir des personnes invités qui sont souvent interviewées. Ce qui explique l'importante proportion de locuteurs non ponctuels dans cette catégorie.
- Les classes les plus représentées notamment dans l'ensemble EPAC-tst sont celles des *journaliste ponctuel* et *autre non ponctuel*.

TABLE 3.2 – Nombre et proportion de chaque rôle dans les ensembles EPAC-app et EPAC-tst.

	présentateur	journaliste non ponct.	autre non ponct.	journaliste ponctuel	autre ponctuel	total
EPAC-app	164(18%)	90(10%)	374(40%)	136(15%)	155(17%)	919(100%)
EPAC-tst	13(12%)	10(9%)	39(36%)	36(35%)	9(8%)	107(100%)

De part les choix qui ont été faits lors de la constitution du corpus EPAC (focus sur la parole conversationnelle et sélection des émissions ou séquences d'émissions à annoter), le corpus EPAC est plus approprié à une étude plus fine sur la reconnaissance des rôles. Le passage à cinq rôles, issu de la distinction entre ponctuels et non ponctuels, que nous proposons s'en trouve d'autant plus justifié.

3.4.2 Mesures d'évaluation

Nous présentons maintenant les différentes métriques que nous avons utilisées pour évaluer nos systèmes sous différents angles.

3.4.2.1 Matrice de confusion

Une matrice de confusion dont un exemple est proposé dans la table 3.3, est un outil de visualisation des résultats d'une tâche de classification [Duda 00]. La matrice rassemble les erreurs de prédiction et permet d'identifier les classes entre lesquelles la confusion est importante. Ce type de matrice peut conduire à remettre en cause certaines classes. Nous utiliserons ces matrices pour appuyer nos analyses.

TABLE 3.3 – Exemple d'une matrice de confusion entre deux classes.

		Classes prédites	
		Classe 1	Classe 2
Classes Réelles	Classe 1	**Correctement Classé**	**Mal Classé**
	Classe 2	**Mal Classé**	**Correctement Classé**

3.4.2.2 Taux de reconnaissance correcte et intervalle de confiance asscocié

Le Taux de Reconnaissance Correcte (TRC) mesure la proportion d'observations correctement classées par rapport au nombre total d'observations à classer :

$$TRC = \frac{\text{Nombre d'observations correctement classées}}{\text{Nombre total d'observations à classer}}$$

Cette mesure est celle utilisée majoritairement dans les travaux de l'état de l'art en reconnaissance automatique du rôle [Barzilay 00, Liu 06, Banerjee 06], mais aussi plus largement en reconnaissance quelle que soit le type de formes et de classes.

Dans toutes nos expériences, nous donnerons l'intervalle de confiance à 95% pour chaque taux de reconnaissance. La largeur de cet intervalle permet de qualifier la qualité des résultats obtenus et de prendre quelques précautions lors de l'interprétation du taux de reconnaissance (TRC). Nous utilisons le calcul suivant : si m désigne le nombre d'observations à classer et $\mathbf{x}$ un vecteur de m valeurs binaires, avec $x(i) = 1$ si la $i^{ième}$ observation est correctement classée, et $x(i) = 0$ dans le cas contraire alors l'intervalle de confiance à 95% se calcule de la manière suivante, avec $\sigma(x)$ l'écart-type et $\overline{x}$ la valeur moyenne :

$$I_{95\%} = \left[\overline{x} \pm \frac{\sigma(x)}{\sqrt{m}}\right]$$

3.4.2.3 Durée de parole traitée correctement classée

Dans des travaux plus récents [Vinciarelli 07, Salamin 09], les performances sont rapportées en terme de durée traitée correctement classée en rôle. Pour pouvoir comparer notre méthode à ces travaux, nous calculons la proportion du temps de parole bien classé τ :

$$\tau = \frac{\text{durée correctement classée}}{\text{durée totale à classer}}$$

3.5 Évaluation de la reconnaissance automatique de rôle - Système à trois rôles

Cette section concerne l'évaluation de plusieurs variantes de notre système de reconnaissance automatique des rôles des intervenants. Comme présenté dans la section 3.1, celui-ci est fondé sur l'architecture classique d'un système de reconnaissance des formes (cf. figure 3.1) et les

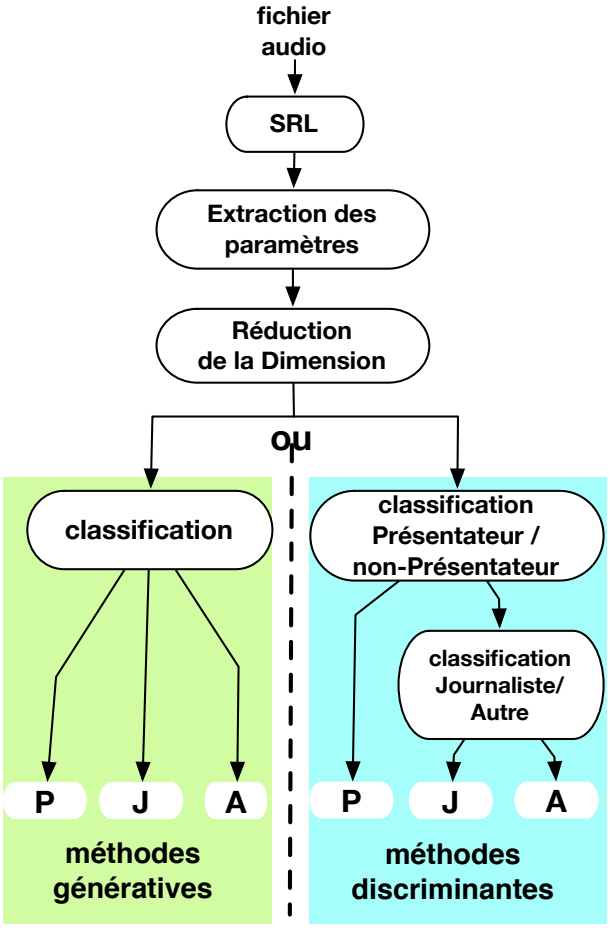


FIGURE 3.6 – Système de reconnaissance automatique basé sur trois rôles et variantes mises en œuvre suivant la méthode de classification utilisée (générative/discriminante).

variantes sont obtenues en combinant différentes méthodes de réduction de dimension avec divers classifieurs.

Dans cette section, nous rapportons trois expériences où nous considérons comme classes les trois types classiques de rôle, à savoir *présentateur*, *journaliste* et *autre*. Pour chacune d'elles, nous évaluons les six variantes de notre système, liées aux six méthodes de classification envisagées (GMM, k-ppv, 4 SVM différents). Suivant la méthode de classification utilisée, la mise en œuvre du système de reconnaissance à trois rôles se fait de deux façons différentes, liant méthode et stratégie adoptée :

- en utilisant une méthode générative (GMM, k-ppv) et une stratégie multi-classes avec les trois classes de rôle {P,J,A} comme représenté par la branche de gauche en couleur verte sur la figure 3.6,
- ou en utilisant une méthode discriminante (SVM linéaire, à noyau RBF, sigmoïdal ou polynomial) et une stratégie qui conduit à l'utilisation de deux classifieurs successifs. Le premier distingue les locuteurs de type *présentateur* des autres intervenants. Ces derniers sont à

leur tour séparés en locuteurs de type *journaliste* ou de type *autre*. Ceci est représenté par la branche de droite, en bleu sur la figure 3.6.

Ce cadre nous permet d'étudier plus particulièrement :

- l'influence des erreurs introduites par l'utilisation d'une segmentation et d'un regroupement en locuteurs automatique en pré-traitement (section 3.5.1),
- l'influence de chacun des sous-ensembles de paramètres acoustiques, temporels et prosodiques considérés individuellement (section 3.5.2),
- l'influence de plusieurs méthodes de réduction de dimension sur les performances de la classification (section 3.5.3).

Dans le but d'alléger la lecture de ce chapitre, nous ne présenterons, dans chaque cas, que les tableaux de résultats servant de base à nos commentaires. Le lecteur intéressé par des résultats complémentaires sera invité à consulter l'annexe A du manuscrit.

3.5.1 Étude de l'influence des erreurs de SRL

Objectifs de l'expérience : à travers cette expérience, nous évaluons l'impact des erreurs de segmentation et de regroupement en locuteurs sur la reconnaissance automatique du rôle en comparant les résultats obtenus à partir d'un étiquetage manuel avec ceux obtenus automatiquement par le biais d'un outil de SRL.

Mise en œuvre du système utilisé : cette première expérience est réalisée avec le corpus ESTER2. La phase d'apprentissage est la même pour toutes les configurations : elle est réalisée en mode supervisé avec l'étiquetage manuel en locuteurs et en rôles. En phase de test, une expérience de reconnaissance est faite à partir de la segmentation manuelle en locuteurs et une autre avec la segmentation automatique. Nous rappelons que l'évaluation de la SRL automatique sur les données d'ESTER-tst atteint DER de l'ordre de 9%. Dans cette expérience, nous n'appliquons pour le moment aucune méthode de réduction de dimension. Notons que compte tenu du faible nombre de données d'apprentissage pour la classe *présentateur*, les GMM sont réduits à une seule loi normale (Modèle de Gaussienne).

Discussion des résultats : les scores obtenus sont regroupés dans la table 3.4. Conformément à ce que nous pouvions supposer, nous observons une baisse entre les performances obtenues à l'aide de la segmentation manuelle et celles obtenues avec la segmentation automatique. Dans le premier cas, les valeurs du taux de reconnaissance vont de 71% à $86,6\%$, alors qu'en version automatique, les valeurs vont de $63,1\%$ à $79,3\%$. Ce phénomène est observé quelle que soit la méthode de classification utilisée : les valeurs baissent en moyenne de 8% avec l'utilisation du prétraitement automatique. Nous observons aussi que globalement la meilleure reconnaissance est obtenue à l'aide des classifieurs SVM. Toutefois cette différence doit être relativisée compte tenu de l'intervalle de confiance à 95% assez large obtenu sur ces expériences : entre $\pm 4,5\%$ et $\pm 6,1\%$ pour le manuel, et entre $\pm 5,9\%$ et $\pm 6,7\%$ pour l'automatique.

Conclusion de l'expérience : cette expérience montre une certaine robustesse de la reconnaissance vis-à-vis des erreurs introduites par la segmentation automatique en locuteurs. Nous observons une baisse d'environ 8% du taux de reconnaissance entre les deux expériences, alors

TABLE 3.4 – Performances du système à 3 rôles, sans réduction de dimension et appliqué (1) sur la segmentation manuelle et (2) sur la segmentation automatique du corpus ESTER-tst.

	$TRC(\%) \pm I_{95\%}$	
	(1) SRL manuelle	(2) SRL automatique
Modèle de Gaussienne	$80,2 \pm 5,3$	$72,4 \pm 6,2$
k-ppv	$71,0 \pm 6,1$	$63,1 \pm 6,7$
SVM rbf	$85,7 \pm 4,7$	$\mathbf{79,3 \pm 5,6}$
SVM polynomial	$\mathbf{86,6 \pm 4,5}$	$79,3 \pm 5,6$
SVM sigmoïdal	$83,9 \pm 4,9$	$75,9 \pm 5,9$
SVM linéaire	$83,9 \pm 4,9$	$75,9 \pm 5,9$

que le traitement automatique a introduit lui-même un pourcentage équivalent d'erreur de segmentation (DER de 9% sur le corpus de test). Ce résultat conforte notre choix d'utiliser les résultats de la SRL automatique dans un tel système, ce qui sera systématiquement fait pour la suite de nos expérimentations. Les lecteurs intéressés retrouveront dans la partie A.1 de l'annexe des expériences complémentaires intégrant également l'utilisation de plusieurs méthodes de réduction de dimension. Il est important de noter que celles-ci ne remettent pas en cause cette conclusion.

3.5.2 Étude de l'influence des sous-ensembles de paramètres

Objectifs de l'expérience : avec cette seconde expérience, nous étudions l'impact de chaque sous-ensemble de paramètres (temporels, acoustiques et prosodiques) sur la reconnaissance des rôles des intervenants.

Présentation du système utilisé : l'ensemble des variantes de notre système, dont l'architecture a été détaillée précédemment, a été testé. Cette expérience est réalisée à l'aide du corpus d'ESTER2 avec en entrée soit :

- le sous-ensemble de paramètres acoustiques,
- le sous-ensemble de paramètres temporels,
- le sous-ensemble de paramètres prosodiques.

Discussion des résultats : les résultats de cette expérience sont rassemblés dans la table 3.5. Le sous-ensemble de paramètres acoustiques, considéré seul, donne des valeurs de TRC entre 35% et $54,7\%$. Il s'agit du sous-ensemble le moins performant de tous. Le sous-ensemble de paramètres temporels permet d'obtenir un taux de reconnaissance bien meilleur entre $67,5\%$ et $70,9\%$. De même, le sous-ensemble de paramètres prosodiques obtient un score allant de $63,5\%$ à $76,4\%$. Ce dernier sous-ensemble permet d'atteindre le taux le plus élevé.

Ces expériences nous mènent à formuler un certain nombre d'observations notamment :

- les paramètres acoustiques ne sont pas très discriminants dans cette éxpérience,

– le sous-ensemble de paramètres prosodiques comme celui des paramètres temporels discriminent bien les *présentateur* du reste des locuteurs, cela est légèrement atténué lorsqu'on les associe,
– les paramètres prosodiques réduisent la confusion entre les classes *autre* et *présentateur*.

Les matrices de confusion, appuyant ces observations, se trouvent en annexe A.2 (matrices A.4, A.6, A.8, A.10). Les tables A.3, A.5, A.7 et A.9 rassemblent des résultats supplémentaires pour lesquels cette même expérience est réalisée, en intégrant une étape de réduction de dimension.

TABLE 3.5 – Performances du système à 3 rôles sur le corpus ESTER-tst, sans reduction de dimension et appliqué sur les paramètres (1) acoustiques, (2) temporels et (3) prosodiques.

	$TRC(\%) \pm I_{95\%}$		
	(1) acoustiques	(2) temporels	(3) prosodiques
Modèle de Gaussienne	$35,0 \pm 6,6$	$70,4 \pm 6,3$	$63,1 \pm 6,6$
k-ppv	$52,7 \pm 6,9$	$69,5 \pm 6,4$	$65,5 \pm 6,5$
SVM rbf	$50,7 \pm 6,9$	$67,5 \pm 6,5$	$75,9 \pm 5,9$
SVM polynomial	$42,8 \pm 7,4$	$70,4 \pm 6,3$	$\mathbf{76,4 \pm 5,8}$
SVM sigmoïdal	$53,7 \pm 6,9$	$\mathbf{70,9 \pm 6,3}$	$73,4 \pm 6,1$
SVM linéaire	$\mathbf{54,7 \pm 6,9}$	$69,5 \pm 6,4$	$71,9 \pm 6,2$

Conclusion de l'expérience : un classement des trois sous-ensembles de paramètres peut être établi. Le sous-ensemble de paramètres acoustiques est le moins performant, puis viennent les paramètres temporels, enfin les paramètres prosodiques qui semblent plus efficaces notamment en réduisant la confusion entre les catégories *autre* et les catégories *journaliste*. Ce classement est indépendant de la méthode de classification utilisée.

Les performances atteintes par un sous-ensemble seul sont cependant inférieures aux scores obtenus grâce à l'ensemble des paramètres temporels et prosodique ou encore à la prise en compte de tous les paramètres. Bien que les résultats sans les paramètres acoustiques soient un peu plus élévés, ce qui n'est pas significatif compte-tenu de l'intervalle de confiance, nous conservons l'ensemble des paramètres dans les investigations à venir, quitte à appliquer automatiquement une méthode de réduction de dimension, comme nous allons le faire dans l'expérience suivante.

3.5.3 Étude de l'influence des méthodes de réduction de dimension

Objectifs de l'expérience : l'expérience précédente montre que les paramètres des trois catégories apportent une information pertinente. Cependant, certains de ces paramètres sont certainement corrélés ou redondants. C'est pourquoi nous tentons d'en réduire le nombre en utilisant les méthodes de réduction de dimension qui visent à « nettoyer » les données des paramètres peu significatifs ou redondants. Nous étudions ici l'influence de plusieurs méthodes de réduction sur la classification.

Présentation du système utilisé : sur le même principe que précédemment, nous testons l'ensemble des variantes du système à trois rôles (figure 3.6) sur le corpus ESTER2. L'étape de réduction de dimension correspond ici à l'application d'une ACP ou d'une AFD. Nous comparons les résultats obtenus à l'aide de ces méthodes à ceux obtenus sans réduction de dimension (voir section 3.5.1) rappelés en colonne (1) de la table 3.6. Nous les comparons également à ceux obtenus avec le sous-ensemble de paramètres ayant donné les performances les plus élevées, c'est-à-dire l'ensemble de paramètres prosodiques (voir section 3.5.2). Des expériences complémentaires sur l'influence des méthodes de réduction de dimension sont proposées en annexe dans les sections A.1 et A.2.

Discussion des résultats : les performances sont présentées dans la table 3.6. L'application de l'ACP, donnant un résultat optimal tout en conservant 99% de l'information initiale, réduit la dimension des données à 20 (contre 36 dans l'espace initial). Dans ce cas, le taux de reconnaissance va de $64,5\%$ à $77,8\%$.

La dimension obtenue par l'application de l'AFD est obligatoirement de 2, puisque le nombre de classes considérées est égal à 3. La classification atteint des résultats allant de $74,4\%$ à $79,8\%$ (TRC). Les paramètres retenus, car maximisant la séparation des classes sur l'ensemble d'apprentissage, sont : $\overline{I}_{seg}$, $\overline{L}_{seg}$, $min(L_{seg})$, $min(P_{env})$, $max(P_{loc})$ et $\overline{P}$ (voir table 2.2 pour la signification de ces paramètres), ce qui correspond à 3 trois paramètres temporels et 3 acoustiques.

Rappelons que le sous-ensemble de 12 paramètres prosodiques conduit à des performances allant de $63,1\%$ à $76,4\%$. Les résultats avec réduction de dimension sont légèrement supérieurs, mais globalement, nous n'observons pas de gain statistiquement significatif.

TABLE 3.6 – Performances du système à 3 rôles sur le corpus ESTER-tst, suivant le nombre ou les combinaisons de paramètres : (1) sans réduction de dimension, ou avec (2) ACP ou (3) AFD, ou (4) avec les paramètres prosodiques seuls.

	$TRC(\%) \pm I_{95\%}$			
	(1) espace initial	(2) après ACP	(3) après AFD	(4) param. pros.
nb de dimensions	36	20	2	12
Mod. de Gauss.	$72,4 \pm 6,2$	$72,4 \pm 6,2$	$78,8 \pm 5,6$	$63,1 \pm 6,6$
k-ppv	$63,1 \pm 6,7$	$64,5 \pm 6,6$	$74,4 \pm 6,0$	$65,5 \pm 6,5$
SVM rbf	$\mathbf{79,3 \pm 5,6}$	$\mathbf{77,8 \pm 5,7}$	$\mathbf{79,8 \pm 5,5}$	$75,9 \pm 5,9$
SVM polynomial	$79,3 \pm 5,6$	$71,4 \pm 6,2$	$73,4 \pm 6,1$	$\mathbf{76,4 \pm 5,8}$
SVM sigmoïdal	$75,9 \pm 5,9$	$74,9 \pm 6,0$	$78,8 \pm 5,6$	$73,4 \pm 6,1$
SVM linéaire	$75,9 \pm 5,9$	$76,4 \pm 5,9$	$79,3 \pm 5,6$	$71,9 \pm 6,2$

Conclusion de l'expérience : au final, les différences de performances relevées entre chacune des méthodes de réduction de dimension ne sont pas statistiquement significatives, compte-tenu de l'intervalle de confiance obtenu. Elles ne permettent pas de sélectionner une méthode en particulier. Nous observons cependant que l'utilisation de l'AFD améliore un peu les performances tout en réduisant considérablement la dimension des données, qui passe de 36 à 2. Le gain obtenu

avec l'AFD doit être de nouveau relativisé compte tenu de l'intervalle de confiance à 95% moyen de $\pm 5,5\%$ qui accompagne ces résultats. Les différentes configurations que nous avons testées dans cette section confortent notre choix :

- de baser la reconnaissance de rôles sur un pré-traitement automatique (SRL),
- de travailler avec un jeu complet de 36 paramètres.

Nous avons mené une autre série d'expériences pour étudier le comportement du système face à une catégorisation un peu plus fine des rôles.

3.6 Évaluation de la reconnaissance automatique de rôle - Système à cinq rôles

Au cours du chapitre 2, nous avons mis en évidence l'importance dans la classification des intervenants qui n'apparaissent qu'une fois, qu'ils soient « journalistes » ou « autres ». Nous introduisons cette distinction entre les locuteurs ponctuels et les locuteurs non ponctuels en faisant évoluer notre système vers un système de reconnaissance de cinq rôles. Dans cette section, nous évaluons progressivement :

- l'impact du passage de trois rôles vers cinq rôles (section 3.6.1),
- l'impact d'une évolution de l'architecture initiale vers une architecture hiérarchique que nous définissons dans la section 3.6.2,
- l'influence du corpus. Nous évaluons notre système de reconnaissance de cinq rôles sur deux ensembles de documents différents : le corpus ESTER2 et le corpus EPAC (section 3.6.3).

3.6.1 Étude de l'influence de la définition des cinq rôles

Objectifs de l'expérience : nous souhaitons étudier l'impact du passage à cinq rôles. Nous rappelons que ces cinq types de rôles recherchés sont : *présentateur*, *journaliste non ponctuel*, *autre non ponctuel*, *journaliste ponctuel* et *autre ponctuel* et qu'un intervenant ponctuel est un locuteur qui n'apparaît que sur un seul segment de parole.

Présentation du système utilisé : la distinction entre les locuteurs ponctuels et les locuteurs non ponctuels appelle à modifier l'architecture du système. Nous présentons sur la figure 3.7 les différentes variantes du système de reconnaissance à cinq rôles. Comme pour le système à trois rôles, ces variantes sont fonction du type de méthode de classification utilisé : génératif ou discriminant. La différence principale avec « l'approche trois rôles » réside dans l'étape de séparation des locuteurs ponctuels et des locuteurs non ponctuels, une fois la segmentation en locuteurs effectuée.

La classification en rôles des intervenants ponctuels réduit le problème de classification à deux classes. Seuls les locuteurs *journaliste* et *autre* peuvent faire partie des locuteurs ponctuels. Ce problème de classification binaire peut donc être traité indifféremment par les méthodes de classification génératives ou discriminantes. Par contre, la classification des locuteurs non ponctuels reste un problème à trois classes. Dans ce cas, nous avons repris les différentes variantes du système à trois rôles déjà vues au cours du paragraphe précédent, à ceci près que les apprentissages

des méthodes se limitent maintenant aux données relatives aux intervenants non ponctuels. Nous observerons l'impact de ces modifications sur les résultats.

Pour cette étude, nous n'utilisons pas dans un premier temps, de méthode de réduction de dimension et nous évaluons toutes les méthodes de classification déjà étudiées dans la section précédente en les appliquant toujours sur les données d'ESTER2 et sur la segmentation automatique en locuteurs.

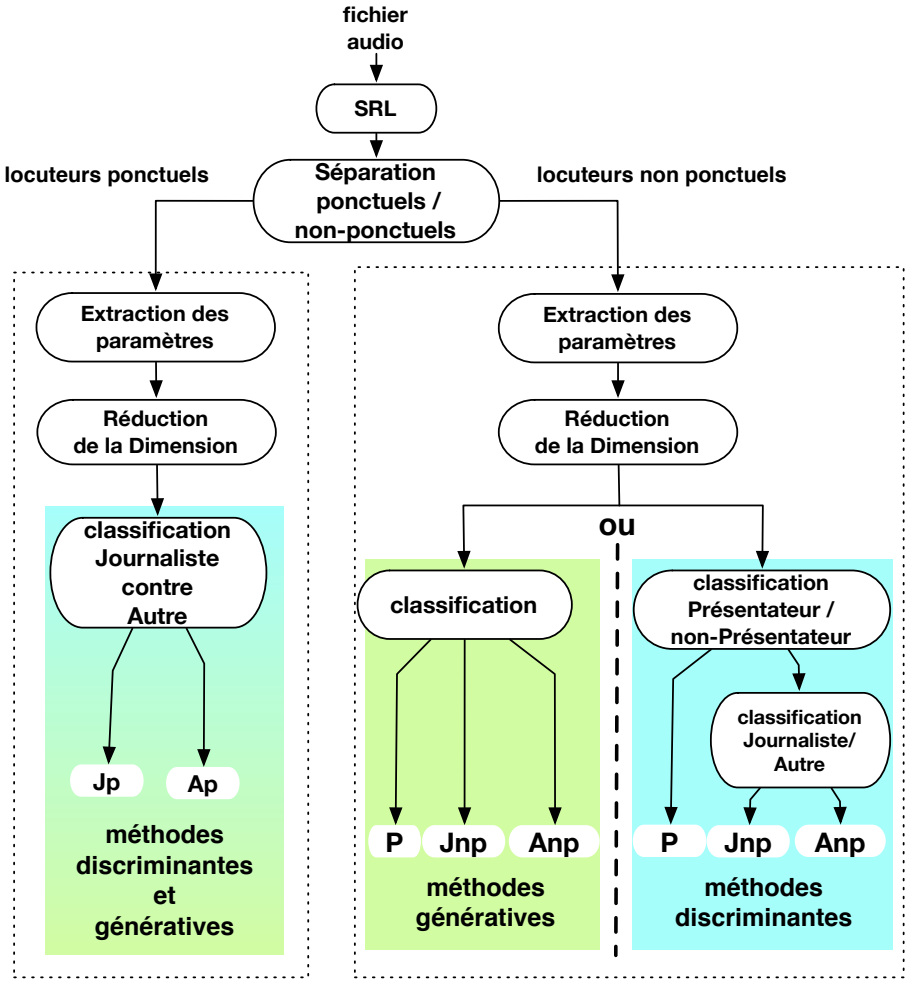


FIGURE 3.7 – Système de reconnaissance automatique basé sur cinq rôles et variantes mises en œuvre suivant la méthode de classification utilisée (générative/discriminante).

Discussion des résultats : la table 3.7 permet de comparer les résultats obtenus avec le système de reconnaissance à cinq rôles et ceux obtenus avec le système à trois rôles. Le système à cinq rôles atteint des performances allant de $61,1\%$ à $81,8\%$ (TRC) tandis que le système à trois rôles permet d'aller de $63,1\%$ et $79,3\%$ (TRC). Ceci représente un gain de $2,5\%$ entre les deux meilleures variantes. Ces différences sont inférieures à l'intervalle de confiance, s'étendant globalement de $\pm 5,4\%$ à $\pm 6,7\%$. Le taux de reconnaissance atteint $81,8\%$, pour le meilleur

système à cinq rôles, utilisant des classifieurs SVM sigmoïdaux. Ces performances sont meilleures que pour le système à trois rôles équivalent (TRC de $75,9\%$).

TABLE 3.7 – Performances des systèmes à (1) trois rôles et (2) à cinq rôles, sans réduction de dimension évalués sur ESTER-tst.

	$TRC(\%) \pm I_{95\%}$	
	(1) 3 rôles	(2) 5 rôles
Modèle de Gaussienne	$72,4 \pm 6,2$	$75,4 \pm 5,9$
k-ppv	$63,1 \pm 6,7$	$61,1 \pm 6,5$
SVM rbf	$\mathbf{79,3 \pm 5,6}$	$77,3 \pm 5,8$
SVM polynomial	$79,3 \pm 5,6$	$77,3 \pm 5,8$
SVM sigmoïdal	$75,9 \pm 5,9$	$\mathbf{81,8 \pm 5,4}$
SVM linéaire	$75,9 \pm 5,9$	$78,8 \pm 5,6$

La table 3.8 présente le détail des scores obtenus sur chaque branche du meilleur système à 5 rôles. Le TRC des locuteurs ponctuels est très bon et atteint $87,3\%$. Le TRC des locuteurs non ponctuels est moins élevé, avec tout de même $79,3\%$ des rôles bien attribués.

TABLE 3.8 – Performances détaillées du système à 5 rôles (1) ponctuels, (2) non ponctuels et (3) total, pour la variante SVM sigmoïdal, sans application de méthode de réduction de paramètres. Les tests sont effectués sur ESTER-tst.

	classification	reduction	nb paramètres	TRC
(1) ponctuels	SVM sigmoïdal	-	24	$87,3\%$
(2) non ponctuels	SVM sigmoïdal	-	36	$79,3\%$
(3) ponctuels + non ponctuels			$TRC(\%) \pm I_{95\%} = 81,8\% \pm 5,4$	

Conclusion de l'expérience : la distinction des ponctuels et des non ponctuels réduit d'office le nombre de paramètres des ponctuels de 36 à 25 paramètres, puisque certains paramètres temporels deviennent redondants ou nuls pour cette catégorie de rôles. Cette séparation réduit aussi considérablement le nombre de données d'apprentissage de chacune des classes.

Malgré cela, les taux de reconnaissance obtenus sont maintenus lors du passage de trois à cinq rôles. Le meilleur système à cinq rôles, correspondant à l'utilisation d'un SVM sigmoïdal qui atteint le taux de reconnaissance le plus élevé soit $81,8\%$, représente même un gain de $2,5\%$ par rapport au meilleur système à trois rôles (qui atteint $79,3\%$ avec un SVM rbf).

3.6.2 Étude de l'influence du passage à une architecture hiérarchique

Objectifs de l'expérience : nous proposons de faire évoluer l'architecture des systèmes précédents vers une architecture hiérarchique dans laquelle toute étape de classification est ramenée à un problème à deux classes, précédée éventuellement d'une étape de réduction de dimension.

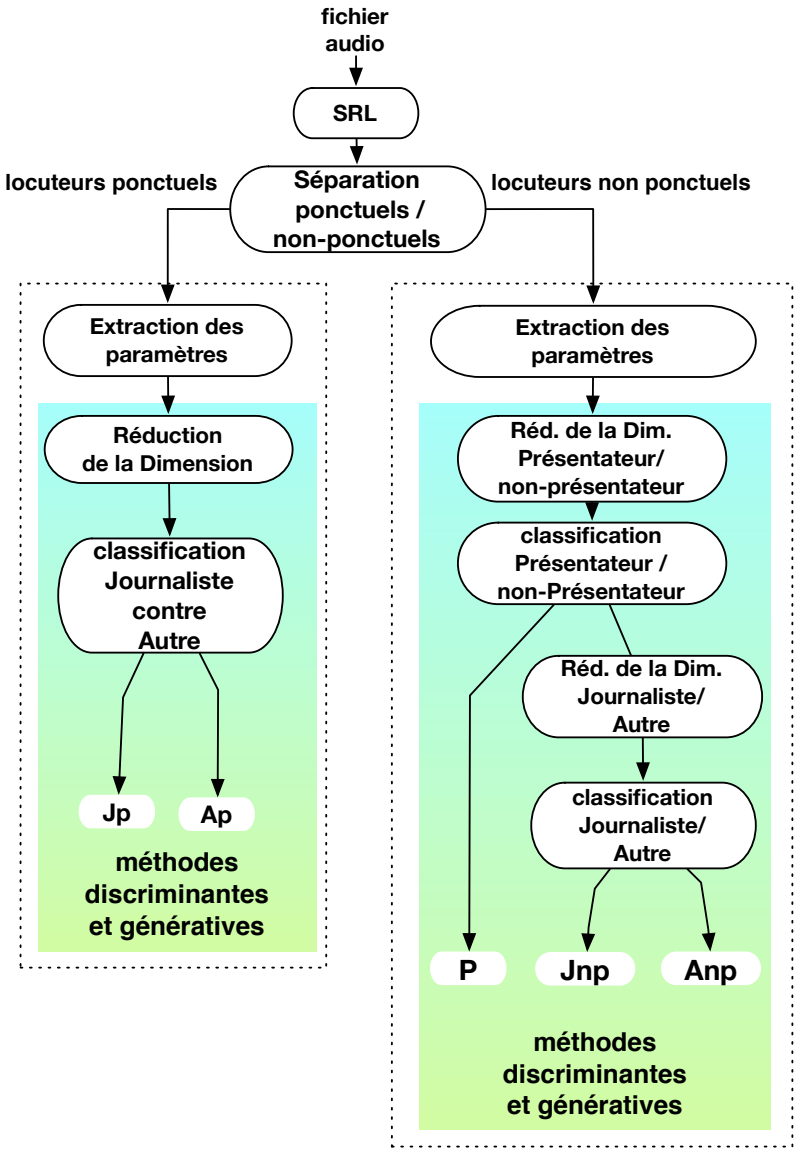


FIGURE 3.8 – Architecture hiérarchique du système de reconnaissance à cinq rôles.

Présentation du système utilisé : l'architecture hiérarchique est présentée sur la figure 3.8. La modification apportée, par rapport à l'architecture précédente (figure 3.7), concerne seulement la mise en œuvre des méthodes génératives (la branche de droite). Il s'agit d'une généralisation de la classification des locuteurs non ponctuels en deux étapes successives. Une première étape de classification va considérer les locuteurs *présentateur* pour les opposer à ceux qui n'en sont pas et désignés par la suite comme *non-présentateur*. Dans une seconde étape, ces derniers sont classés en *journaliste non ponctuel* et *autre non ponctuel*. Les classifications en deux classes étaient auparavant réalisées uniquement avec des SVM : nous étendons cette approche à toutes les méthodes de classification utilisées. L'architecture hiérarchique intègre l'application d'une méthode de réduction de dimension, par transformation ou par sélection, avant chaque étape de

classification. Dans cette expérience, nous utilisons deux méthodes de classification génératives (Modèle de Gaussienne et k-ppv), et deux classifieurs discriminants(SVM linéaire et à noyau rbf). Nous appliquerons systématiquement une sélection de paramètres par éliminations successives (RSE), avant l'étape de classification.

L'intérêt principal du système hiérarchique est de permettre une analyse des paramètres discriminants à chaque étape de classification ramenée à deux classes. Les performances ne seront rapportées que dans le but de montrer que cette architecture ne dégrade pas la reconnaissance. Nous passerons plus de temps à décrire les paramètres retenus à l'étape de réduction de la dimension. Nous reportons dans la table 3.9, les performances obtenues pour le cas non hiérarchique (colonne (1)), et pour le cas hiérarchique (colonne (2)). Les détails d'autres expériences menées avec l'architecture hiérarchique sont donnés dans la section A.4 de l'annexe A.

TABLE 3.9 – Performances obtenues sur ESTER-tst, par les variantes du système à cinq rôles dans sa version (1) non hiérarchique sans réduction de dimension ou (2) hiérarchique avec application de la méthode de sélection de paramètres RSE.

	$TRC(\%) \pm I_{95\%}$	
	(1) non hiérarchique	(2) hiérarchique
Modèle de Gaussienne	$75,4 \pm 5,9$	$75,8 \pm 5,9$
k-ppv	$61,1 \pm 6,5$	$64,5 \pm 6,6$
SVM rbf	$77,3 \pm 5,8$	$72,4 \pm 6,2$
SVM linéaire	$\mathbf{78,8 \pm 5,6}$	$\mathbf{81,3 \pm 5,4}$

Discussion des résultats : les meilleures performances atteintes dans les deux cas par la variante SVM linéaire, sans réduction de dimension, s'étendent de $61,1\%$ à $78,8\%$ (TRC) pour le système non hiérarchique et de $64,5\%$ et $81,3\%$ (TRC) pour le système hiérarchique.

Nous rapportons maintenant une description des paramètres conservés par le meilleur système, intégrant un classifieur SVM linéaire (le détail est donné dans la section A.4 de l'annexe A) :

- à l'étape de classification des *présentateur*, la RSE conserve 14 paramètres, dont 5 temporels, 3 acoustiques et 6 prosodiques, pour un score élevé de 92% (TRC),
- à l'étape de classification *journaliste non ponctuel* contre *autre non ponctuel*, elle conserve 26 paramètres pour un taux de reconnaissance de 74,6% (TRC) : 14 paramètres temporels, 4 acoustiques et 8 prosodiques,
- enfin, à l'étape de classification des rôles *journaliste ponctuel* et *autre ponctuel*, la RSE conserve 9 paramètres pour obtenir un score de 88,9% (TRC) : 3 paramètres acoustiques et 6 prosodiques.

Il est intéressant de noter que dans tout ces différents cas, quatre paramètres sont systématiquement conservés. Il s'agit de :

- $var(P_{loc})$: la variance de la puissance du signal sur les zones attribuées au locuteur,
- $max(F_0)$: la valeur maximale du pitch,
- N_{sil} : le nombre de silences,
- $\overline{Duree_{sil}}$: la durée moyenne des silences.

Conclusion de l'expérience : les performances du système hiérarchique à cinq rôles dans sa variante SVM linéaire, couplée avec la méthode de sélection de paramètres, donne un score de $81,3\%$ avec un intervalle de confiance de $\pm 5,4\%$. Les performances restent du même ordre que celles observées dans les expériences précédentes. Plusieurs choses intéressantes sont mises en évidence par cette expérience, notamment :

- la très bonne classification des *présentateur*,
- le caractère discriminant des paramètres temporels pour les locuteurs non ponctuels,
- les paramètres acoustiques et prosodiques sont conservés à chaque étape de classification du meilleur système : ceci laisse à penser que ces descripteurs sont particulièrement pertinents.

Avec la mise à disposition du corpus EPAC en 2010, nous avons pu réaliser une série d'expériences, complémentaires mais non exhaustives grâce auxquelles nous avons pu étudier l'influence du corpus sur le système de reconnaissance hiérarchique à cinq rôles.

3.6.3 Étude de l'influence du corpus

Objectifs de l'expérience : afin d'étudier la robustesse face au corpus d'apprentissage et de test, le système hiérarchique à cinq rôles, évalué sur le corpus ESTER dans les précédentes sections, est maintenant évalué sur le corpus EPAC.

Présentation des systèmes utilisés : dans les deux cas, nous utiliserons l'architecture hiérarchique à cinq rôles avec sélection de paramètres (RSE) et le classifieur SVM linéaire. Nous rappelons que :

- le corpus EPAC se compose de EPAC-app (80 heures) et de EPAC-tst (10h), respectivement les ensembles d'apprentissage et de test,
- le corpus ESTER2 se compose ESTER-dev (6 heures) et ESTER-tst (7 heures).

Les résultats correspondants sont reportés dans la table 3.10.

TABLE 3.10 – Meilleures performances du système hiérarchique à cinq rôles avec sélection de paramètres (RSE) en fonction du corpus.

App / Test	$TRC(\%) \pm I_{95\%}$
ESTER-dev / ESTER-tst	$81,3 \pm 5,4$
EPAC-app / EPAC-tst	$83,2\% \pm 7,12$

Discussion des résultats : sur le corpus ESTER2, la reconnaissance réalise un score de $81,3\%$. Sur EPAC, le TRC est de $83,2\%$ (TRC). L'intervalle de confiance important relativise ce gain. Nous rapportons, pour le corpus EPAC, la liste des paramètres conservés au moment de la sélection de paramètres de chaque étape de classification. Nous constatons que :

- à l'étape de classification des locuteurs de type *présentateur*, la RSE conserve 17 paramètres, dont 3 paramètres temporels, 8 acoustiques et 6 prosodiques,

– à l'étape de classification *journaliste non ponctuel* contre *autre non ponctuel*, elle conserve 14 paramètres : 5 paramètres temporels, 5 acoustiques et 4 prosodiques,
– enfin, à l'étape de classification des rôles *journaliste ponctuel* et *autre ponctuel*, la RSE conserve 11 paramètres : 1 paramètre temporel, 4 acoustiques et 6 prosodiques.

La nature des paramètres conservés est détaillée en annexe (section A.5.3).
Plusieurs éléments intéressants méritent d'être soulignés :
– sur les deux corpus, les paramètres temporels sont discriminants, particulièrement pour les locuteurs non ponctuels,
– contrairement au corpus ESTER2, les paramètres acoustiques semblent plus discriminants sur le corpus EPAC, notamment dans la classification *présentateur* contre *non présentateur*.

Ici aussi, quatre paramètres sont systématiquement conservés à chaque étape de classification. Il s'agit de deux paramètres acoustiques, et de deux paramètres prosodiques :
– $\overline{P_{env}}$: puissance moyenne du signal sur les zones de non parole,
– $min(P_{loc})$: valeur minimale de la puissance du signal sur les zones de parole,
– $Debit_{sil}$: nombre de silences par unité de temps,
– $var(Duree_{sil})$: variance de la durée des silences.

Comme pour le cas du corpus ESTER2, plusieurs paramètres prosodiques sont conservés. Cela semble confirmer l'importance des informations prosodiques dans la caractérisation des rôles. La matrice de confusion associée à cette étude est disponible en annexe dans la table A.20. Nous y constatons que les locuteurs *autre ponctuel* sont très bien classés. La confusion est plus importante entre *présentateur* et *autre non ponctuel*.

Conclusion de l'expérience : les différences observées entre les deux corpus ne se situent pas au niveau des scores obtenus mais plus particulièrement au niveau des paramètres conservés pour chaque étape de classification. Notons sur les deux corpus :
– l'importance des paramètres temporels pour les locuteurs non ponctuels en particulier,
– l'importance des paramètres acoustiques et prosodiques. Ils jouent donc un rôle discriminant à toutes les étapes de classification.

Nous avons également étudié un ensemble plus large d'expériences sur EPAC autour de plusieurs variantes de systèmes.

3.7 Expérience sur le corpus EPAC

Objectifs de l'expérience : nous évaluons à travers une série d'applications bien ciblées, les performances atteintes par plusieurs systèmes, intégrant le corpus EPAC.

Présentation des systèmes utilisés : nous utilisons l'architecture hiérarchique à cinq rôles. Nous avons testé différentes variantes autour des méthodes de classifications suivantes :

- le SVM Linéaire : ce classifieur a rapporté de bonnes performances lors de expériences précédentes,
- les GMM : la taille du corpus EPAC étant plus importante, nous avons pu utiliser cette méthode.

Discussion des résultats : des résultats détaillés sont reportées dans la table 3.11. Nous remarquons que :

- (1) le système hiérarchique à cinq rôles, sans réduction de dimension, couplé avec une classification par GMM, rapporte un taux de reconnaissance de $74,8\% \pm 8,26$. Le nombre de composantes Gaussiennes utilisées sont de 2 pour le *présentateur*, 8 pour les *journaliste non ponctel* et *autre non ponctuel*, et de 4 pour les ponctuels. Ce score est finalement décevant : nous nous concentrerons dans la suite sur l'utilisation du classifieur SVM.
- (2) la variante fondée sur un SVM linéaire sans réduction de dimension, rapporte un taux plus élevé, de l'ordre de $88,9\% \pm 6,00$. Nous avons également décliné les différentes méthodes de réduction de dimension.
- (3) l'ajout d'une AFD n'améliore pas les résultats, $(88,8\% \pm 6)$ tout en réduisant la dimension des données à une seule dimension.
- (4) l'utilisation d'une sélection de paramètres RSE fait chuter les performances à $83,2\% \pm 7,1$. Les paramètres retenus ont été présentés dans l'expérience précédente.
- (5) l'utilisation de l'ACP permet à la reconnaissance d'atteindre un TRC de $92\% \pm 5,3$. Il s'agit du score le plus important que nous aillons obtenu sur l'ensemble des expériences menées.

TABLE 3.11 – Performances des variantes GMM et SVM linéaire du système hiérarchique à 5 rôles sur le corpus EPAC. Ce système est testé avec différentes méthodes de réduction de dimension.

méthode	réduction	dimensions	$TRC(\%) \pm I_{95\%}$
(1) GMM (2/8/4)	aucune	24/36/36	$74,8\% \pm 8,2$
(2) SVM linéaire	aucune	24/36/36	$88,9\% \pm 6,0$
(3) SVM linéaire	AFD	1/1/1	$88,8\% \pm 6,0$
(4) SVM linéaire	RSE	11/17/14	$83,2\% \pm 7,1$
(5) SVM linéaire	ACP	25/19/19	$92\% \pm 5,3$

Conclusion de l'expérience : sans que nous puissions dire qu'il s'agit du *meilleur système* (du fait de l'intervalle de confiance relativement important qui accompagne l'ensemble de résultats de cette section), l'utilisation conjointe du SVM linéaire et de l'ACP avec l'architecture hiérarchique, permettent de reconnaître 92% des rôles des intervenants, ce qui est un excellent score. Les matrices de confusion, reportées en annexe A.21, permettent d'observer que les quelques erreurs restantes concernent des *autre non ponctuel* classés comme des *présentateur*. Cette confusion s'explique par le fait que le corpus contient quelques exemples d'émissions dans lesquelles la distinction entre invité principal et présentateur n'est pas évidente : l'émission prend

la forme d'une conversation et le temps de parole du présentateur et de l'invité sont équivalents. Dans le chapitre suivant, nous travaillerons avec le corpus EPAC. Nous utiliserons les résultats fournis par ce système.

3.8 Conclusion

Dans ce chapitre, nous avons présenté un ensemble d'expériences réalisées autour de l'étude et du développement d'un système de reconnaissance automatique des rôles des locuteurs avec détection automatique des locuteurs. Notre système s'appuie sur la structure classique d'un système de reconnaissance des formes incluant une phase de pré-traitement des données, une extraction de paramètres, une éventuelle réduction de la dimension des données et une étape de classification.

Nous avons étudié à travers un grand nombre d'expériences, l'influence de plusieurs méthodes de réduction de dimension : analyse en composantes principales (ACP), analyse factorielle discriminante (AFD) et une méthode de sélection de paramètres par Recherche Séquentielle par Élimination(RSE). Nous avons également évalué plusieurs méthodes de classification supervisées : GMM, k-ppv, SVM à noyau linéaire, rbf, sigmoïdal et polynomial. Nous avons réalisé une reconnaissance automatique des rôles des locuteurs sur deux ensembles de documents : le corpus ESTER2 qui contient en majorité des bulletins d'information et le corpus EPAC qui comporte une proportion importante de parole conversationnelle (débats, interviews et magazines).

Les premières expériences ont concerné la reconnaissance automatique des trois rôles classiques, rencontrés dans la littérature du domaine. Les performances de reconnaissance des trois rôles *présentateur*, *journaliste* et *autre* atteignent sur le corpus ESTER2, dans le meilleur cas, $\mathbf{79,3\% \pm 5,6}$. Ce score est obtenu à l'aide d'un système fondé sur l'utilisation d'un classifieur SVM à noyau rbf, sans réduction de la dimension. Ce score est comparable aux performances rapportées dans la littérature : entre 80% et 85% de rôles bien reconnus.

Nous avons ensuite présenté une seconde architecture, de manière à intégrer cinq catégories de rôles. La version la plus aboutie de cette architecture consiste en un système de classification hiérarchique, dont la particularité est de ramener toutes les étapes de classification à des problèmes à deux classes. Cette architecture permet également de faire une étude plus spécifique des paramètres les plus discriminants pour chaque étape de classification.

En pratique, le meilleur score obtenu sur le corpus ESTER2, à l'aide d'un système intégrant l'architecture hiérarchique, un classifieur SVM linéaire et une étape de sélection de paramètres RSE, atteint un TRC de $\mathbf{81,3\% \pm 5,4}$. Le même système appliqué au corpus EPAC rapporte un TRC de $\mathbf{83,2\% \pm 7,1}$. Ce système nous permet de mettre en évidence la place importante tenue par les descripteurs prosodiques dans la reconnaissance des rôles puisque plusieurs descripteurs calculés à partir de la fréquence fondamentale et des silences détectés sont retenus systématiquement. Finalement, le meilleur score de classification est obtenu sur le corpus EPAC, il consiste en une ACP et un classifieur SVM linéaire. Dans ce cas, $\mathbf{92\% \pm 5,3}$ des rôles sont correctement attribués aux locuteurs.

Ces très bons résultats nous permettent d'envisager l'utilisation de ce système de reconnaissance de rôles, dans notre système de structuration des documents audiovisuels, auquel se consacre le chapitre suivant.

Chapitre 4

Structuration de documents audiovisuels et rôles des intervenants

Le système hiérarchique de reconnaissance de rôles, présenté dans le chapitre précédent, a rapporté de très bons résultats avec 92% de rôles correctement détectés et identifiés sur le corpus EPAC. Ce résultat nous permet d'envisager d'intégrer l'information relative aux rôles dans un travail de recherche en structuration par le contenu de documents audiovisuels. Nous présentons, dans ce chapitre, une contribution à ce domaine.

La structure d'un document peut être vue comme une succession de segments temporels, chacun étant homogène et représentatif d'un contenu. Dans notre cas, l'objectif est de faire émerger d'un document audiovisuel non structuré, une représentation structurelle où chaque segment fait référence à un ensemble de locuteurs typés par leur rôle et interagissant selon un certain mode. De ce fait, notre approche exploite les résultats du système hiérarchique de reconnaissance automatique des rôles (chapitre 3), ainsi que ceux du système de détection des zones d'interaction orale (chapitre 1).

La suite du chapitre s'organise de la manière suivante. La section 4.1 présente un état de l'art de la recherche en structuration de documents audiovisuels par le contenu. Dans la section 4.2 nous précisons le cadre dans lequel se situe la structuration proposée. Puis, avant de conclure, nous présentons le système de structuration fondé sur le rôle des intervenants (section 4.3), ainsi que son évaluation réalisée à l'aide du corpus EPAC (section 4.4).

4.1 La recherche en structuration par le contenu de documents audiovisuels

Les approches en structuration par le contenu des documents audiovisuels se répartissent principalement en deux catégories selon le niveau de granularité recherché. La première catégorie de méthodes s'attache à structurer le flux audiovisuel tandis que la seconde catégorie concerne la structuration interne d'un programme ou d'un ensemble de programmes. Les résultats recherchés par chacune de ces catégories sont différents, ce qui implique des approches également très différentes dont nous tentons de donner un aperçu dans la suite de cette section.

4.1.1 Structuration du flux audiovisuel

À l'échelle d'un flux ou d'un document contenant plusieurs heures de contenu audiovisuel, un objectif commun aux méthodes de structuration est de retrouver la suite d'événements relatifs au contenu de la grille des programmes. Il s'agit dans ce cas de structurer le flux en deux types d'événements : les **programmes** et les **inter-programmes**. Un segment est soit une instance d'un programme, soit une inclusion entre de tels programmes.

D'après [Manson 10], dont le travail concerne la délinéarisation des flux télévisuels, les **inter-programmes** se rassemblent en 6 catégories : *publicité, bande-annonce, parrainage, jingle, auto-promotion et campagne d'intérêt général*. Les **programmes** sont définis par le même auteur, comme des événements du flux liés par une charte audiovisuelle, vecteurs d'une valeur culturelle, informative ou divertissante.

Pour retrouver les programmes et inter-programmes dans le flux audiovisuel, il est naturel de vouloir se référer à la grille des programmes (ou EPG [7]) et de l'exploiter quand elle est disponible. Elle correspond à une liste ordonnée et datée des programmes diffusés ou « programmation ». Les horaires de diffusion et les noms des programmes sont souvent accompagnés d'informations complémentaires telles que le genre ou la durée des programmes.

Les travaux récents de [Poli 07, Naturel 07, Manson 10] rappellent la fiabilité toute relative des informations contenues dans l'EPG. Les horaires peuvent être imprécis, ne permettant pas de retrouver avec justesse les bornes de début et de fin des émissions. Dans cette grille, ne sont pas répercutées nécessairement les modifications de programmation induites par des faits d'actualités importants. L'EPG ne contient généralement aucune information (ou très peu) sur les inter-programmes. Pour palier ces limites, les méthodes actuelles s'appuient sur une analyse directe des documents par leur contenu, éventuellement complétée par certaines méta-données issues de la grille des programmes.

Les méthodes de structuration du flux sont fondées, de manière exclusive, soit sur la détection des **programmes**, soit sur la détection des **inter-programmes**. À notre connaissance, il n'existe pas de méthode réalisant la détection simultanée de ces deux événements.

Les **inter-programmes** sont les événements le plus largement exploités dans la littérature. En effet, ils présentent une propriété de redondance intéressante que n'ont pas les programmes. Certains inter-programmes, comme les publicités en particulier, sont rediffusés à de nombreuses reprises et sur plusieurs chaînes de télévision différentes. Plusieurs approches exploitent cette propriété de redondance, soit à travers une recherche directe des répétitions dans le flux audiovisuel [Covell 06], soit par une comparaison du flux avec une base de données contenant des exemples d'inter-programmes [Naturel 07, Pua 04, Wen 99]. Cette catégorie de méthodes implique un traitement supplémentaire pour maintenir l'ensemble de références à jour. D'autres approches détectent indirectement les inter-programmes, notamment à travers la recherche de ruptures dans le flux audiovisuel, comme des images monochromes et des zones de silences [Dimitrova 02, Lienhart 97, Sadlier 02].

Les méthodes fondées sur la détection des **programmes** sont moins nombreuses. La raison principale évoquée dans [Naturel 07] est que « les programmes n'exhibent aucune caractéristique

7. Electronic Program Guide

commune qu'il serait possible de détecter directement ». Dans le manuscrit de [Manson 10] est dressé un panorama récent des différentes approches de structuration de flux audiovisuels. Plusieurs travaux recherchent directement les programmes dans le flux de données en détectant des invariants de production caractéristiques de certains programmes. L'approche de [Liang 05] s'appuie sur la recherche de génériques particuliers de début et de fin de programmes. La méthode de [Wang 08] réalise une détection d'images spécifiques dans le flux vidéo, qui combinées à une analyse audio, permettent de trouver une transition entre un inter-programme et un programme. À l'opposé, la contribution de [El Khoury 08] recherche des ruptures dans l'homogénéité du flux audiovisuel, sans connaissance *a priori*. Une rupture est un changement significatif du contenu audio et vidéo pouvant correspondre à un début ou à une fin de programme. Cette méthode est la seule exploitant les propriétés intrinsèques du contenu d'un programme, pouvant de ce fait être généralisée à plusieurs types de programmes.

4.1.2 Structuration interne d'un programme audiovisuel

Cette seconde catégorie rassemble des méthodes appliquées à l'échelle d'une émission, qu'il faut comprendre ici dans le sens d'une « instance d'un programme ».

L'objectif de ces méthodes est de générer un sommaire de l'émission, définissant un ensemble de zones sémantiquement cohérentes. Ces approches s'appuient généralement sur des descripteurs dits de « bas-niveau », extraits directement du signal audio et vidéo. La difficulté principale rencontrée par ces méthodes est de « franchir le fossé sémantique » séparant les descripteurs « bas-niveau » des unités structurelles d'un plus haut niveau symbolique et permettant de décrire le déroulement d'un programme.

À notre connaissance, il n'existe pas de méthode générique permettant de réaliser cette étape. Les travaux du domaine se restreignent généralement à un seul type de programme. Un grand nombre de contributions concernent plus particulièrement la structuration de vidéos de journaux d'information, de rencontres sportives et de jeux télévisés :

- les journaux télévisés suivent une structure imposée par les règles de production. Ils se composent généralement d'une série de thématiques variant en fonction de l'actualité. Le format de présentation en revanche est redondant d'un bulletin à l'autre. La présentation d'un sujet est souvent composée d'une introduction lue par le présentateur suivie d'un développement dans un reportage. La littérature regorge de travaux exploitant la connaissance *a priori* de cette structure pour segmenter les bulletins d'informations en sujets [Maybury 96, Merlino 97, Hauptmann 98, Kemp 03, Meinedo 03, Chua 04, Ma 09], etc.
- dans les vidéos de sport, les structures temporelles des rencontres sont contraintes par les règles officielles et les méthodes de structuration s'appuient sur la connaissance de ces règles. Les travaux de [Kijak 03] concernent l'analyse et la reconnaissance de la structure de vidéos de rencontres de tennis. Cette contribution établit une modélisation statistique de la structure de vidéos de match par des Modèles de Markov Cachés dont les topologies intègrent les règles du sport.
- les jeux télévisés, également dictés par des règles, sont des programmes très structurés. Les travaux de [Javed 02] et [Ibrahim 07] sont appliqués à des émissions de ce type. Ils fournissent une segmentation en phases de jeu permettant un parcours non linéaire des émissions.

Le travail de [Kolluru 07] va au-delà de la structuration simple en identifiant les segments. Il concerne plus précisément la catégorisation des sujets de bulletins d'information en fonction de leur format de présentation. Ce travail utilise les résultats d'une segmentation et d'un regroupement en locuteurs, d'une segmentation en sujets, et du résultat d'une reconnaissance en entités nommées. Les locuteurs sont classés en trois rôles *présentateur*, *journaliste* et *autre* en fonction de règles simples basées sur l'ordre d'apparition des locuteurs, et de phrases clés extraites de la transcription.

Quatre types de segments, correspondant à quatre formats de présentation des sujets, dépendant des rôles des intervenants, sont proposés. Ces catégories sont :

- une *tribune*, quand le présentateur présente et développe seul le sujet,
- une *correspondance*, quand le sujet est présenté par le présentateur et par un journaliste,
- une *reportage*, quand les trois rôles sont présents,
- une *interview*, pour un sujet concernant le présentateur et un locuteur de la classe *autre*.

Finalement 66% des formats de présentation sont correctement reconnus.

Comme nous allons le développer dans la section suivante, la détection et l'identification des rôles nous permettent d'intervenir sur ces deux facettes de la structuration, la localisation des unités de programmes et la décomposition d'un programme. L'intégration d'un ensemble de règles basées sur les rôles des intervenants nous conduisent à définir des unités de structuration telles que celles présentées par [Kolluru 07].

4.2 Contribution : utilisation des rôles pour la structuration des documents audiovisuels

4.2.1 Positionnement de notre étude

Nous souhaitons exploiter la connaissance du rôle des intervenants pour structurer les documents audiovisuels à l'échelle d'un flux et à l'échelle d'un programme.

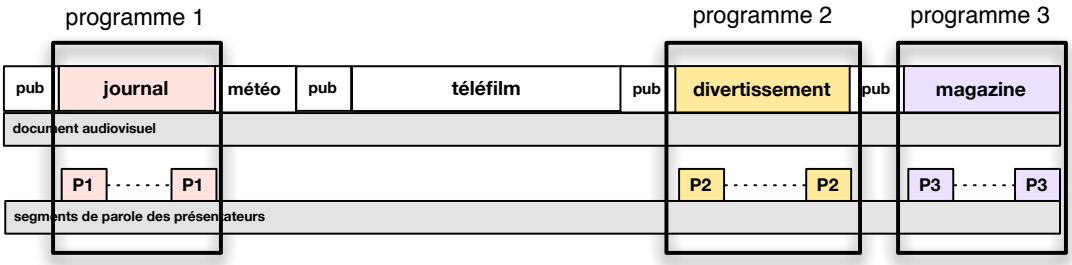


FIGURE 4.1 – Flux audiovisuel composé de plusieurs programmes avec un présentateur par programme.

À l'échelle d'un flux audiovisuel : la connaissance de la présence du présentateur peut être utilisée pour détecter les zones du document correspondant à des instances de programme. Comme nous l'illustrons sur la figure 4.1, intuitivement il semble possible de coupler une zone

temporelle du flux correspondant à une émission avec les informations relatives aux interventions des présentateurs. Cette détection de programmes pourrait être appliquée à tous les types de programmes couverts par les interventions d'un présentateur, comme les journaux d'information, les émissions de divertissement et les magazines culturels.

Plusieurs points motivent l'utilisation des informations relatives au présentateur. Le présentateur est garant du bon déroulement de l'émission, de ce fait ses interventions sont rarement improvisées. Il incombe à ce locuteur de présenter le sommaire de l'émission, d'introduire les autres intervenants (candidats, invités, chroniqueurs), de les remercier et de clore l'émission. De plus, comme nous l'avons observé dans le chapitre précédent, le rôle du présentateur est particulièrement bien reconnu par notre système automatique et sa détection peut servir de point d'ancrage.

Nous rassemblons, dans la figure 4.2, des exemples de programmes (journaux d'information, magazines musicaux, jeux télévisés) dans lesquels les premières et les dernières personnes prenant la parole sont les présentateurs, indiquant ainsi le début et la fin de l'émission.

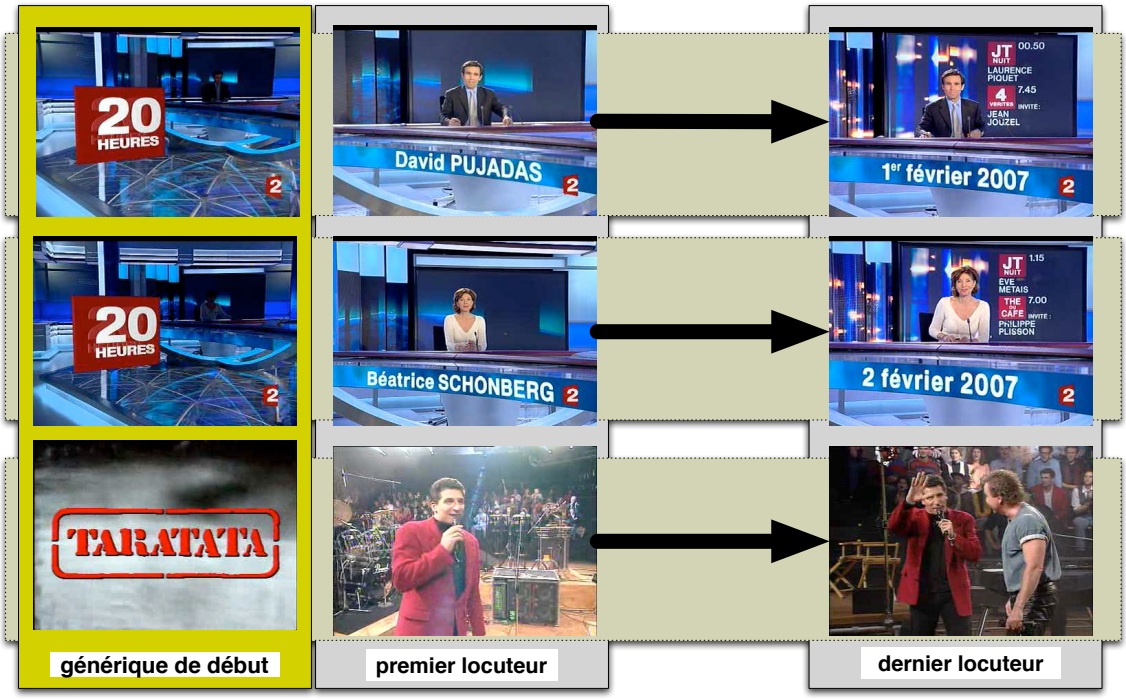


FIGURE 4.2 – Exemples d'émissions où le premier et le dernier mot prononcé est systématiquement attribuable au présentateur.

Par ailleurs, il est fréquent que le public identifie un programme à la personne qui le présente. Toutefois, il faut rester prudent et ne pas faire l'amalgame entre le rôle et l'identité du présentateur. Le journaliste Patrick Poivre d'Arvor (cf. figure 4.3) a présenté le journal de 20 heures de la première chaîne française pendant plus de 20 ans. Il incarne sûrement « le 20h de TF1 » auprès du public, pourtant à d'autres occasions, comme nous pouvons le voir sur la figure 4.3, il pourra être reçu par un confrère, en tant qu'invité cette fois, pour parler par exemple de son actualité littéraire. La relation « présentateur-émission » est bien évidemment mise en défaut

comme le montrent également les deux premiers exemples de la figure 4.2 en illustrant que le rôle de présentateur d'un même programme peut être occupé par deux individus différents, sans pour autant modifier la structure du programme. Pour toutes ces raisons, nous pensons qu'il est préférable de s'appuyer sur l'information « rôles des intervenants », information plus robuste que l'identité des individus, pour la tâche que nous souhaitons réaliser.

FIGURE 4.3 – Exemple d'un présentateur de TF1, invité sur canal+.

Dans plusieurs autres genres de programmes (cf. travail de [Poli 07] pour une liste exhaustive), comme les documentaires et les retransmissions sportives, il est plus difficile d'évoquer la présence d'un présentateur. Toutefois les interventions du narrateur d'un documentaire ou d'un commentateur sportif présentent des caractéristiques assez similaires à celles des présentateurs.

Nous ne traiterons pas de programmes de fiction tels que les séries, les films ou les téléfilms, qui échappent à cette définition et nécessiteraient donc une étude spécifique.

À l'échelle d'un programme : de manière comparable à la contribution de [Kolluru 07], notre méthode vise à exploiter la connaissance des rôles des intervenants pour caractériser le contenu des programmes.

Considérons l'exemple d'une émission quotidienne comme un talk-show. Les invités, les chroniqueurs, les reporters, les personnages publics ou anonymes apparaissant dans ces émissions changent presque chaque jour. Il en est de même pour les candidats d'un jeu télévisé, les artistes invités dans une émission de variétés ou les participants d'un débat ; tous ces individus seront amenés à varier à travers les épisodes d'un programme.

La structuration automatique des programmes a donc *a priori* intérêt à s'appuyer sur les rôles des intervenants, qui représentent une information robuste d'une émission à une autre.

La suite de ce chapitre est consacrée à la description de notre travail. Nous commençons par poser un ensemble de définitions fondamentales à notre approche, avant de décrire notre méthode de structuration automatique et de l'évaluer.

4.2.2 Définitions des éléments de structuration

Nous décrivons ici les unités logiques de structuration qui nous permettent de décrire la structure de documents audiovisuels. Pour ce faire, nous postulons que : « **dans un flux ou un document audiovisuel, une zone temporelle durant laquelle un ou plusieurs locuteurs non-présentateurs interviennent accompagnés par au plus un présentateur**

est une séquence correspondant à une sous partie de programme, éventuellement à un programme complet ».
Nous appellerons une telle zone, *une brique élémentaire de structuration*. Dès lors que cette brique fera intervenir un présentateur, commencera et finira par une intervention de ce présentateur, cette brique sera appelée *brique élémentaire bornée*.

Ce postulat nous conduit à définir deux types d'unités de structuration – les unités « présentées » et les unités « intermèdes » – qui seront fondamentales dans notre approche :

- les **unités « présentées »** sont les briques élémentaires bornées maximales d'un document, dans le sens où cette brique ne peut être contenue dans une brique de même nature. Il s'en suit que ces unités peuvent être de deux types :
  - Majoritairement ce sont des unités faisant intervenir un ou plusieurs locuteurs autour du seul présentateur ; nous parlerons **d'unités couvertes** par un présentateur.
  - Les autres unités se réduisent à l'intervention du seul présentateur. Cette situation peut se produire lors d'un passage de relais entre deux émissions.
- les **unités « intermèdes »** sont les zones complémentaires. Durant ces zones, aucun présentateur n'est présent : il peut s'agir de zones publicitaires, de la présentation du bulletin météo, de plage musicale... Elles peuvent aussi contenir des locuteurs intervenant dans les segments adjacents, auquel cas une partie de ces zones peut être rattachée au segment adjacent concerné.

Ces deux types d'unités sont représentés sous la forme de segments temporels en bas de la figure 4.4. Nous avons cherché à caractériser ces zones plus finement en fonction des rôles et des interactions des locuteurs qu'elles rassemblent.

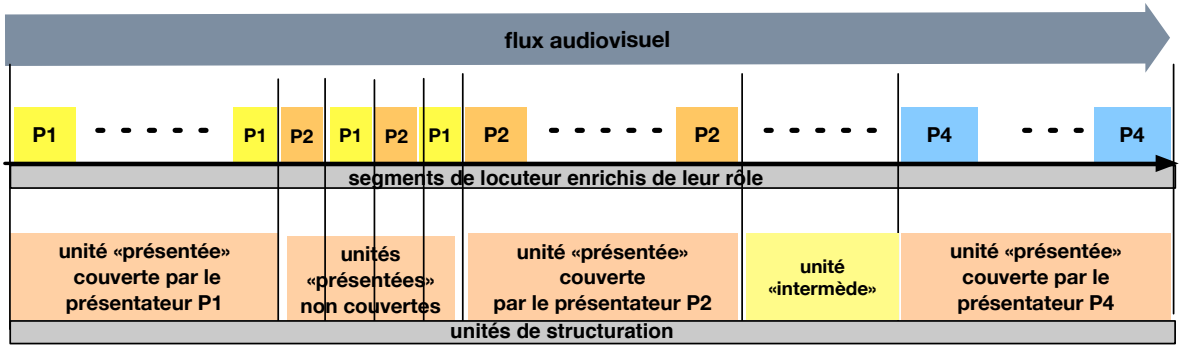


FIGURE 4.4 – Exemples d'unités de structuration.

4.2.3 Les deux niveaux de structuration

La prise en compte d'informations plus précises sur les intervenants au travers de leur rôle et du type d'interaction a donné naissance à deux niveaux de structuration.

4.2.3.1 Premier niveau de structuration : prise en compte du rôle

Le premier niveau de structuration assigne les unités « présentées » dans plusieurs catégories. Nous proposons trois catégories d'unités « présentées » : « informations », « entretiens » et « transition ». La distinction entre ces trois catégories est faite en rapport au temps de parole imputable à chaque type de rôle présent sur cette unité. Nous considérons l'ensemble des locuteurs présents sur l'unité et qui ne sont pas des présentateurs, nous nommerons $\mathbf{d}(\mathbf{journaliste})$ le temps de parole cumulé de tous les journalistes de l'unité, et $\mathbf{d}(\mathbf{autre})$ le temps de parole cumulé des intervenants qui ne sont ni présentateur, ni journaliste.

- Les unités « présentées » de type « informations » correspondent aux unités de programmes sur lesquelles

  $$\mathbf{d}(\mathbf{journaliste}) \geq \mathbf{d}(\mathbf{autre})$$

  Cette catégorie correspond aux unités structurantes pour lesquelles des journalistes sont présents et parlent majoritairement. Dans les journaux, les sujets sont souvent présentés conjointement par le présentateur et un journaliste. Les interventions d'invités, ou d'interviewés, sont moins nombreuses et couvrent moins de temps de parole.
- Les unités « présentées » de type « entretiens » sont les zones durant lesquelles

  $$\mathbf{d}(\mathbf{autre}) > \mathbf{d}(\mathbf{journaliste})$$

  Au moins un intervenant est de type *autre*, donc potentiellement invité, et donc il est souhaitable que dans cette catégorie de programmes les invités occupent la plus grande proportion de parole.
- Les unités « présentées » de type « transition » sont des segments contenant uniquement l'intervention d'un présentateur. Cet type d'inter-programme, plus courant à la radio qu'à la télévision, est caractéristique, comme nous l'avons dit, d'une transition entre deux programmes prenant la forme d'une discussion informelle entre deux présentateurs.

Les unités définies précédemment sont exploitées directement dans le second niveau de structuration.

4.2.3.2 Second niveau de structuration : prise en compte du type d'interaction

Des interactions orales entre intervenants peuvent avoir lieu dans une zone du document attribuée à l'une des catégories « informations » ou « entretiens ». En fonction du contexte et des rôles qui y sont impliqués, cette interaction peut correspondre à des événements différents. Il peut s'agir par exemple de l'interview d'un invité, menée par un présentateur durant un journal télévisé, ou bien d'une discussion entre deux invités, durant un débat de société. Nous attribuons les zones d'interaction (z.i.) découvertes, parmi quatre catégories en fonction des règles suivantes :

- le type « interview » est attribué dans notre méthode à une z.i. entre un présentateur et un interviewé,
- le type « chronique » est attribué à une z.i. impliquant un présentateur et un journaliste,
- le type « débat » correspond à une interaction entre deux invités, ou entre un invité et un journaliste.
- le type « relais » est attribué aux z.i. entre deux présentateurs.

4.2.3.3 Résumé

La figure 4.5 reprend les définitions des différentes unités de structuration recherchées ainsi que leur hiérarchies. Dans la suite de ce chapitre nous décrivons notre algorithme de détection des unités « présentées » et « intermèdes », ainsi que le typage des interactions.

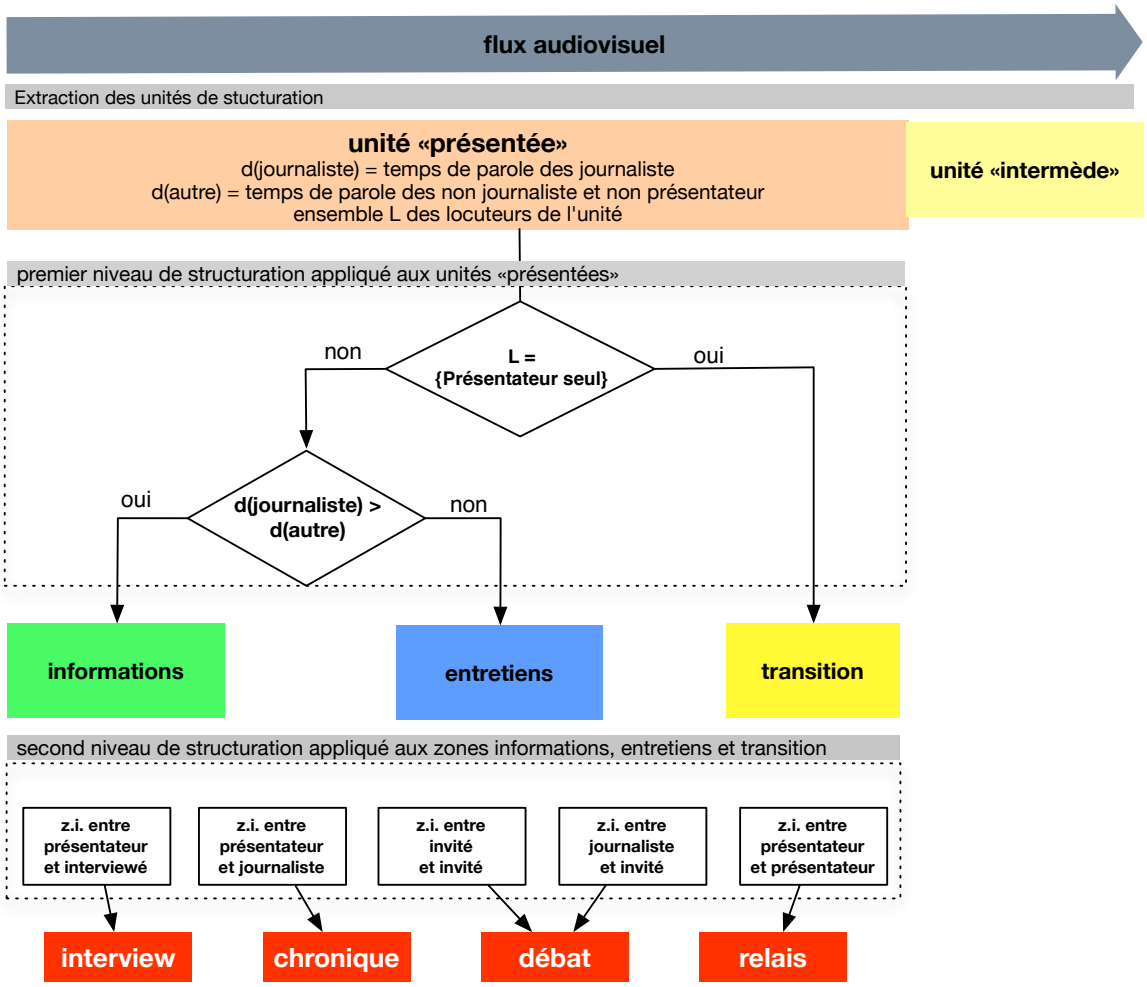


FIGURE 4.5 – Méthode de structuration à partir de la connaissance des locuteurs, de leur rôle et leur type d'interaction.

4.3 Présentation du système de structuration automatique

Le système automatique développé vise à rendre compte de la structuration définie au paragraphe précédent, dans le but de l'étendre à un corpus conséquent et d'évaluer ainsi cette proposition. En particulier, il s'agit de mesurer si les erreurs de reconnaissance (des rôles et des zones d'interaction) ont un impact fort sur la structuration finale du document, telle qu'elle a été définie au paragraphe précédent.

Ce système se décompose en trois modules principaux, présentés sur la figure 4.6 :

– un pré-traitement, où l'on retrouve le processus de segmentation en locuteurs, le système de reconnaissance automatique des rôles et le système de recherche des interactions entre locuteurs,
– un module visant à rendre compte d'une structuration de premier niveau au travers d'une macro-segmentation,
– un module délivrant la structuration finale.

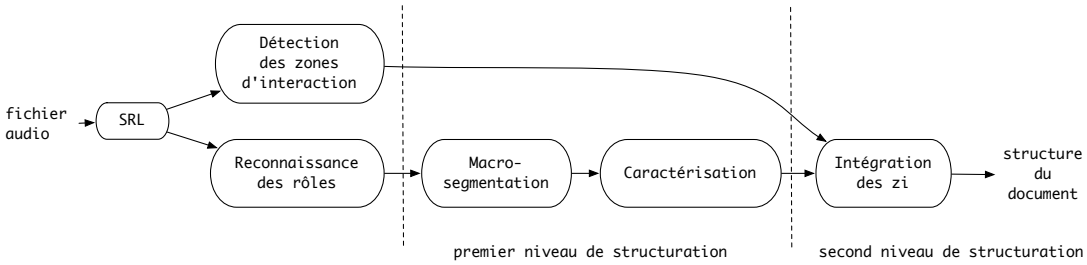


FIGURE 4.6 – Notre système de structuration de contenus audio, utilisant la connaissance des rôles de locuteurs et des zones d'interaction orale.

Rappelons que les systèmes de détection des zones d'interaction (décrit dans le chapitre 1) et de reconnaissance automatique des rôles (décrit dans le chapitre 3) produisent respectivement :

– une segmentation en zones d'interaction. Les tours de parole des locuteurs impliqués dans une zone d'interaction ne doivent pas être séparés par un intervalle sans parole plus long qu'une seconde.
– l'attribution d'un rôle parmi cinq aux locuteurs détectés automatiquement. Cette catégorisation est faite en utilisant les rôles *présentateur*, *journaliste ponctuel* et *journaliste non ponctuel*, *autre ponctuel* et *autre non ponctuel*, reconnus par le système hiérarchique.

Ces deux systèmes utilisent en entrée le résultat de la SRL automatique (décrite dans le chapitre 1). Ces éléments sont présentés dans la partie de gauche de la figure 4.6.

Pour réaliser l'extraction des unités « présentées » et « intermèdes », une macro-segmentation est générée, fondée sur l'hypothèse de « couverture par un présentateur ».

Pour compléter le premier niveau de structuration, les macro-segments sont ensuite classés en « entretiens », « informations », « transition » ou « intermèdes » à partir de l'information sur les rôles. Le rôle *journaliste*, évoqué précédemment dans la description de la méthode, rassemble les rôles *journaliste ponctuel* et *journaliste non ponctuel*, trouvés automatiquement. Les rôles « interviewé », « invité » et « autre » correspondent aux rôles automatiques *autre ponctuel* et *autre non ponctuel* trouvés par le système de reconnaissance automatique de rôles.

Dans un second niveau de structuration, les zones d'interaction détectées appartenant à des unités « présentées » sont enrichies en fonction des rôles des locuteurs impliqués dans l'interaction. Dans la suite de ce paragraphe, nous présentons le processus de macro-segmentation et les deux étapes de classification.

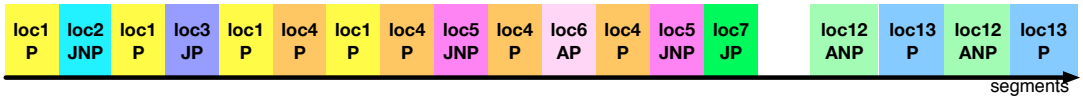


FIGURE 4.7 – Résultat d'une segmentation en locuteurs enrichie par les rôles.

4.3.1 Méthode de macro-segmentation fondée sur les rôles des locuteurs

La macro-segmentation a pour but de retrouver les éléments de base de la structuration, que sont les unités « présentées » et les « intermèdes ».

En entrée de ce traitement, nous avons un ensemble L de N locuteurs. Chaque locuteur est caractérisé par son rôle r_i, avec $r_i \in \{P, JNP, JP, ANP, AP\}$, et ainsi $L = \{loc_{(1,r_1)}, \ldots, loc_{(N,r_N)}\}$.

Chaque locuteur $loc_{(i,r_i)}$ est représenté par un ensemble de segments temporels définis individuellement par leurs instants de début et de fin. Nous illustrerons ce résultat grâce à l'exemple de la figure 4.7. Sur cette représentation, chaque locuteur est indiqué par une couleur unique, par un identifiant loc_i, ainsi que par le symbole de son rôle. Notre algorithme de macro-segmentation se décompose en trois étapes de traitement.

• **Détection du début d'un macro-segment :** cette étape recherche les instants où un présentateur laisse sa place à un autre présentateur. L'algorithme parcourt les segments dans l'ordre chronologique. Dès qu'un autre présentateur apparaît, l'algorithme marque une frontière D_j au début de l'intervention du « nouveau » présentateur ; ce présentateur est le $j^{ième}$ présentateur intervenant depuis le début du traitement. À l'initialisation, comme il n'y a pas de présentateur « courant », l'algorithme indique donc une frontière D_1 au début du premier segment de présentateur rencontré. Au terme de cette étape, la segmentation initiale comporte M frontières D_j avec $j = 1 \ldots M$ associée chacune à un présentateur $loc_{(i,P)}$ de l'ensemble L. La figure 4.8 illustre ce résultat, pour l'exemple choisi en figure 4.7.

• **Détection de fin d'intervention d'un présentateur :** pour chaque frontière de début $\mathbf{D_j}$, l'algorithme cherche une frontière F_j, correspondant à la fin de la dernière intervention du

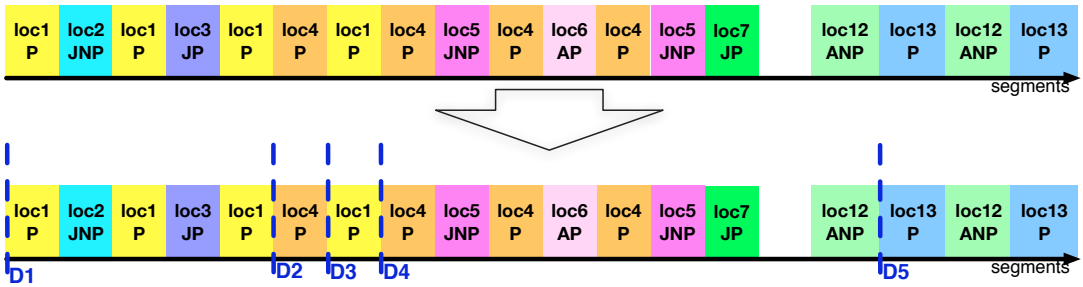


FIGURE 4.8 – Macro-segmentation fondée sur les rôles : détection des instants où un nouveau présentateur apparaît.

présentateur associé à D_j. L'illustration est donnée sur la figure 4.9. Les frontières F_j y sont indiquées en couleur rouge.

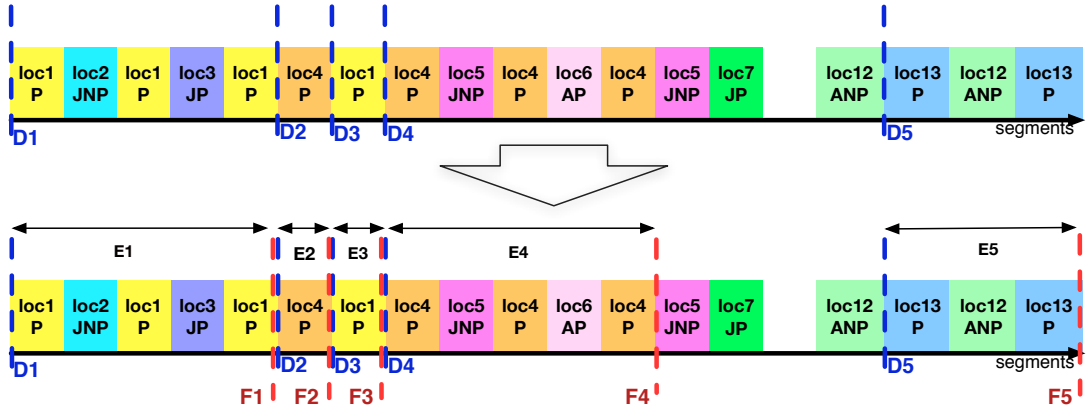


FIGURE 4.9 – Macro-segmentation fondée sur les rôles : détection des frontières délimitant les fins de couverture des présentateurs pour chaque zone Ej.

• **La macro-segmentation finale :** à l'issue de ces deux premières étapes, les segments $E_j = [D_j, Fj]$, $j = 1 \ldots M$ correspondent à des unités « présentées ». Néanmoins, il nous apparaît normal de rattacher à ces unités les segments de parole extérieurs à cette unité (adjacents et correspondants à des locuteurs présents dans l'unité « présentée »). Compte tenu de leur rôle (journaliste ou autre), ces locuteurs interviennent nécessairement dans une même unité de programme. La troisième et dernière étape de l'algorithme de segmentation va tenter d'affiner les bornes des segments pour rendre compte de cet état.

Nous prenons en considération les locuteurs non-présentateurs présents sur l'événement E_j. Nous nommons cet ensemble de locuteurs $L(E_j)$. Durant cette étape, les bornes D_j et F_j vont être déplacées pour prendre en compte les segments de locuteurs de $L(E_j)$ se trouvant aux frontières extérieures de la séquence et les y intégrer.

L'algorithme teste l'identifiant du locuteur apparaissant juste avant la frontière D_j. Si ce locuteur appartient à $L(E_j)$, alors la frontière D_j est décalée pour intégrer ce segment dans E_j. Cette opération est répétée jusqu'à ce que la condition précédente ne soit plus vraie. La même opération est réalisée à la frontière de fin de séquence F_j. Sur la figure 4.10 nous indiquons en vert les frontières qui ont été réajustées lors de cette étape. L'ensemble de ces frontières définit la macro-segmentation finale.

4.3.2 Classification des macro-segments

Cette méthode de macro-segmentation fait émerger trois types de segments (cf. figure 4.11) :

– la première catégorie de macro-segments correspond aux événements E_j contenant l'intervention d'un présentateur seul, sans aucun autre locuteur. Nous les nommons **P-seul**,

– le second type de macro-segments correspond aux événements E_j contenant un présentateur ainsi que d'autres intervenants de rôle différents. Nous les nommons **P+autres**,

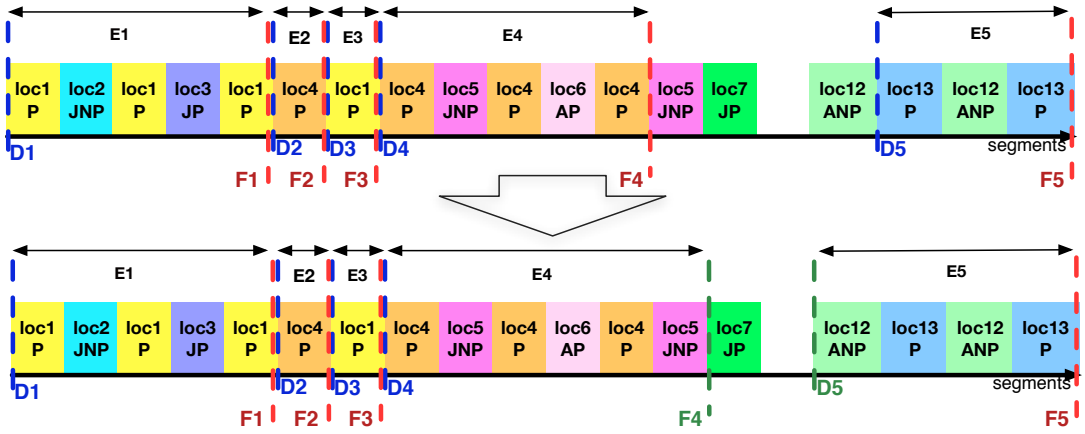


FIGURE 4.10 – Macro-segmentation fondée sur les rôles : affinage des frontières de chaque zone Ej.

– la troisième catégorie correspond aux séquences de segments de locuteurs ne contenant aucun présentateur : ils sont de type $[F_j, D_{j+1}]$. Nous les nommons **sans-P**.

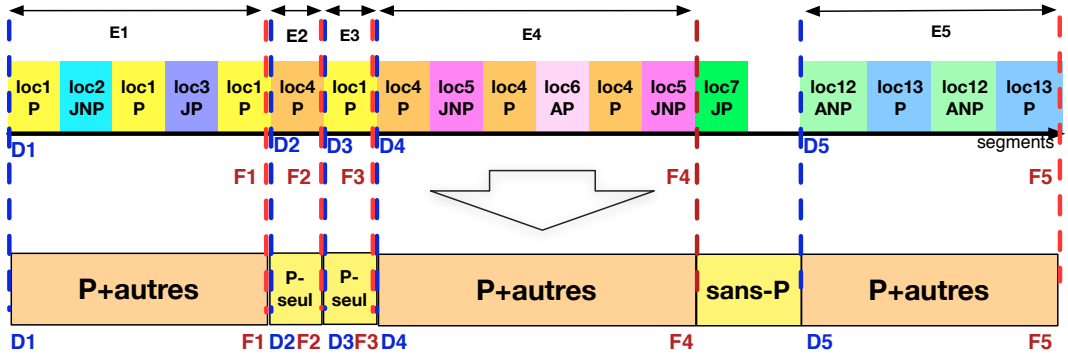


FIGURE 4.11 – Macro-segmentation fondée sur les rôles : les types de macro-segments obtenus.

Il faut remarquer que les segments obtenus ne sont pas exactement les unités « présentées » et « intermèdes » du fait de l'ajustement des frontières. Néanmoins, les règles adoptées au paragraphe précédent pour définir les notions « informations », « entretiens » et « transition » restent applicables. Il s'en suit la catégorisation immédiate suivante :

– une unité « présentée » est obtenue à partir des macro-segments de type **P+autres**. Les étiquettes « entretiens » et « informations » sont attribuées en fonction des temps de parole cumulés attribuables à chaque type de rôles.
– les unités de type « transition » correspondent aux macro-segments de type **P-seul**. Si plusieurs segments adjacents sont étiquétés « transition », ils sont regroupés en un seul segment « transition ».
– les unités « intermèdes » correspondent aux macro-segments de type **sans-P**.

La figure 4.12 illustre cette ultime étape.

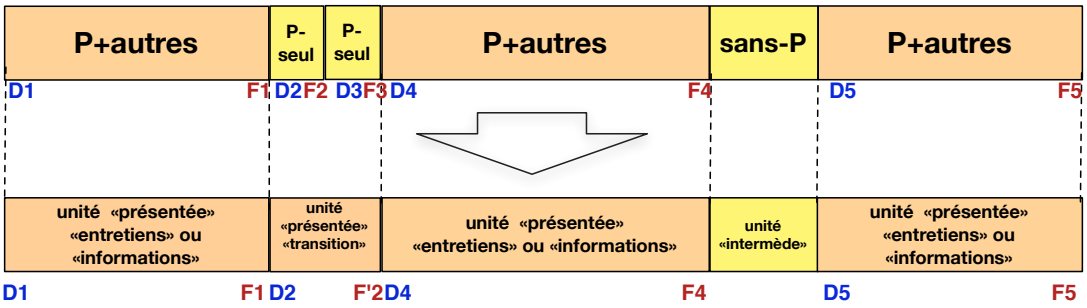


FIGURE 4.12 – Macro-segmentation fondée sur les rôles : résultat du premier niveau de structuration automatique.

4.3.3 Catégorisation des zones d'interaction

Les zones d'interaction détectées qui sont incluses temporellement dans des unités « présentées », sont caractérisées durant cette étape. En fonction des rôles des locuteurs impliqués dans l'interaction, la zone est attribuée à l'une des quatre catégories suivantes : « interview », « chronique », « débat », « relais », comme nous l'avons détaillé plus tôt dans la figure 4.5. Au contraire, les zones d'interaction incluses temporellement entre les instants de début et de fin d'une unité « intermèdes » ne sont pas considérées. La figure 4.13 illustre cette étape de la structuration. Sur cet exemple, toutes les z.i. sont catégorisées car elles coïncident temporellement avec une ou plusieurs unités « présentées ».

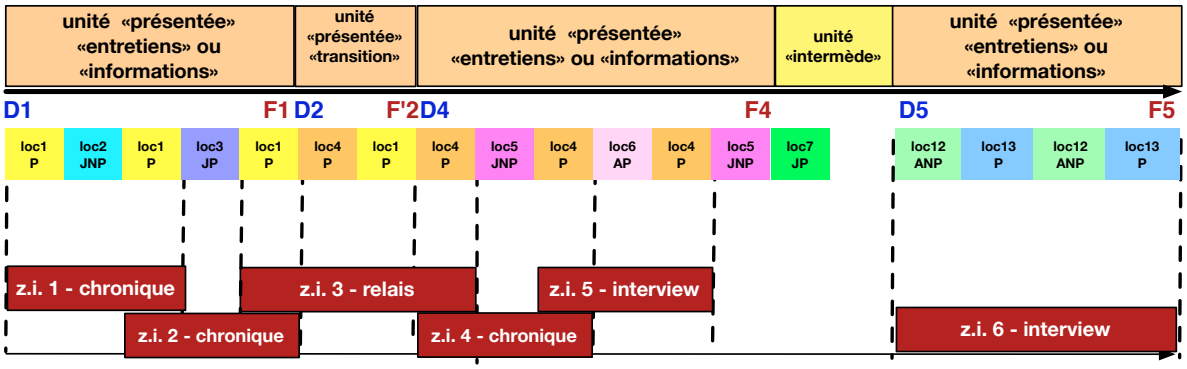


FIGURE 4.13 – Catégorisation des zones d'interaction (en rouge), à l'aide de la connaissance des bornes temporelles des unités « présentées » et « intermèdes » (en haut) et de la connaissance des rôles des locuteurs impliqués dans ces zones d'interaction (au milieu).

4.4 Évaluation de la structuration fondée « locuteurs »

La structuration théorique proposée est formulée à partir de données exactes en termes de nombre de locuteurs, de rôles des locuteurs et en termes d'interaction, sans garantie de résultats en pratique. C'est pourquoi il est nécessaire d'évaluer d'une part la robustesse de cette

proposition face aux erreurs inévitables produites par l'automatisation du traitement, et d'autre part la généricité de l'approche en traitant plusieurs types d'émissions.

4.4.1 Le protocole expérimental

L'évaluation est réalisée à l'aide des documents de l'ensemble de test du corpus EPAC. Ce corpus, que nous avons présenté dans la section 3.4.1.2, rassemble 10 heures d'enregistrements radiophoniques francophones, de types divers. Néanmoins, certaines émissions n'ont pas été annotées manuellement dans leur totalité : seules les séquences de parole dite conversationnelle l'ont été systématiquement. Pour évaluer correctement notre méthode de structuration, il est souhaitable que le corps des émissions ne soit pas tronqué. Nous avons donc complété le corpus afin que cette contrainte soit satisfaite et nous avons ajouté manuellement les annotations pour les besoins de cette expérience. L'annotation réalisée concerne les informations relatives aux tours de parole et aux rôles des intervenants. Finalement, l'ensemble d'évaluation dans l'expérience rassemble 10h45 d'audio.

Nous comparons deux structurations du corpus :

- la structuration dite de référence (vérité terrain), obtenue en appliquant le système de structuration automatique aux résultats d'un étiquetage manuel en locuteurs et rôles (déjà utilisé auparavant pour évaluer notre système hiérarchique de reconnaissance en rôles),
- la structuration automatique, obtenue comme son nom l'indique, par le système totalement automatique.

L'évaluation concerne le premier et le second niveau de structuration. Les différents types d'unités « présentées » et « intermèdes », obtenus par un traitement totalement automatique, sont confrontés à la vérité terrain, ainsi que les étiquetages des zones d'interaction. Les comparaisons sont faites quantitativement (pourcentage de durée de document bien classée) et qualitativement sur quelques exemples types d'émissions.

4.4.2 Évaluation quantitative des deux niveaux de structuration

Dans cette première expérience, nous évaluons le résultat de **la structuration de premier niveau** à partir des segments de types :

- « informations », « entretiens », « transition » pour les unités « présentées »,
- « intermèdes » pour les autres unités.

Les performances de la structuration automatique sont exprimées comme le rapport τ, entre la durée de document correctement étiquetée et la durée de document traitée (une définition est présentée dans le chapitre 3) :

$$\tau = \frac{\text{durée correctement classée}}{\text{durée totale}}$$

La reconnaissance atteint un score τ égal à 85,6%. Nous rapportons, à travers la table 4.1, les confusions entre les quatre classes recherchées. Les valeurs sont exprimées d'une part en secondes et d'autre part en pourcentages.

L'origine de la confusion observée entre les différentes classes découle naturellement des erreurs de SRL et de reconnaissances des rôles. En particulier, la dispersion importante dans des

TABLE 4.1 – Matrice de confusion entre les unités de la vérité terrain et les unités obtenues automatiquement en secondes et en pourcentages.

		traitement automatique			
		informations	entretiens	transition	intermèdes
Vérité terrain	informations	5759 (66,8%)	2187 (25,3%)	0 (0%)	678 (7,8%)
	entretiens	1066 (3,6%)	26769 (91,3%)	407 (1,4%)	1078 (3,7%)
	transition	16 (16%)	6 (6%)	53 (53%)	25 (25%)
	intermèdes	12 (1,8%)	64 (9,9%)	8 (1,2%)	563 (87%)

zones de types « entretiens » est liée aux erreurs de regroupement commises par la SRL lorsque le présentateur d'une émission parle en même temps que le générique de début. Les confusions entre les unités « entretiens » et « informations » sont causées par les confusions entre les rôles *autre* et *journaliste* introduites par le système de reconnaissance des rôles.

La seconde expérience concerne l'évaluation du **second niveau de structuration** du système. Plus précisément, les zones d'interaction, détectées automatiquement et catégorisées selon les règles rassemblées dans la figure 4.5, sont comparées à la vérité terrain obtenue dans les conditions énoncées plus tôt. Les catégories de zones d'interaction sont au nombre de quatre : « chronique », « débat », « interview » et « relais ».

Comme nous l'avions expliqué dans le chapitre 1, deux zones d'interaction adjacentes peuvent se recouvrir sur la durée d'un segment. De ce fait, la vérité terrain et le résultat de la structuration automatique peuvent contenir des segments auxquels correspondent deux catégories.

La comparaison du résultat du traitement automatique avec la vérité terrain révèle que 67,1% de la durée des zones d'interaction détectées automatiquement appartiennent exactement aux mêmes types d'événements.

Les erreurs observées s'expliquent par le fait que les rôles *journaliste* et *autre*, qu'ils soient ponctuels ou non, influent beaucoup plus dans cette étape de caractérisation. En effet, le système de reconnaissance automatique de rôles étant un petit peu moins performant sur ces catégories de rôles, il est normal de constater une confusion un peu plus importante à cette étape. Le second point pouvant expliquer les erreurs entre vérité terrain et résultats automatiques tient du fait que les zones d'interaction détectées sont extrêmement sensibles aux erreurs de segmentation locales introduites par la SRL.

4.4.3 Discussion autour de quelques exemples

Trois types d'extraits sont étudiés : un passage contenant plusieurs programmes, une émission de type débat et une émission de type matinale.

4.4.3.1 Structuration d'un document contenant plusieurs programmes

Nous observons les résultats de structuration obtenus sur un enregistrement de 75 minutes, issu de la station de radio RFI (tranche horaire de 16h55 à 18h10). Ce document rassemble un magazine culturel, un jingle, un bulletin d'information, une auto-promotion et un second magazine culturel.

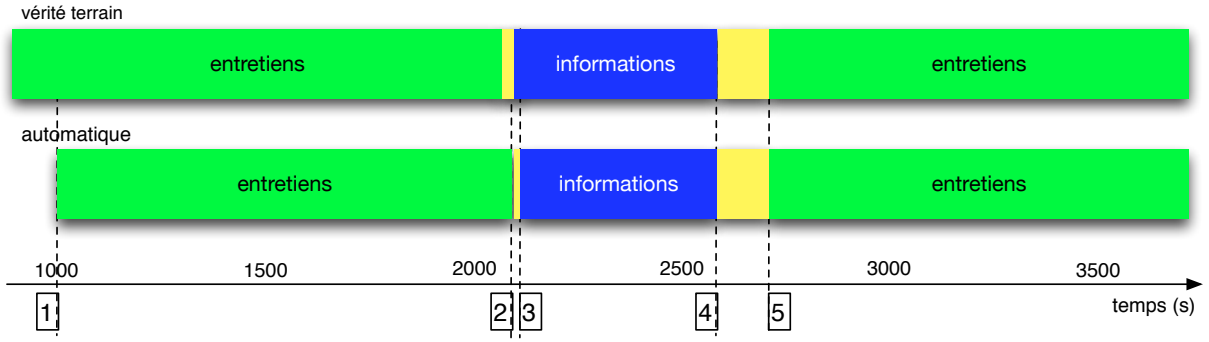


FIGURE 4.14 – Résultats du premier niveau de structuration sur une tranche horaire contenant plusieurs programmes. La vérité terrain est représentée sur le frise du haut, le résultat automatique se trouve sur la frise du bas. Les « entretiens » sont indiqués en vert, les « informations » en bleu et les « intermèdes » en jaune.

Premier niveau de structuration. Nous représentons, sur la figure 4.14, les bornes temporelles des événements de la vérité terrain (en haut) et ceux obtenus automatiquement (en bas). La séquence d'unités de programmes a été retrouvée : aucun événement n'est omis lors de la détection automatique. De plus, les types d'unité « entretiens » et « informations » ont été correctement attribués.

Afin de faciliter l'analyse des résultats, nous commentons des instants importants, repérés sur la figure 4.14 par un numéro.

1 : Le document débute par une séquence d'une centaine de secondes pendant lesquelles le présentateur parle sur un fond musical. Cette zone n'est pas détectée par le système dès l'étape de SRL. En effet, le regroupement des zones de parole superposée à de la musique est un point faible de la méthode utilisée. L'unité de programme débute dès que la musique cesse.

2 : La frontière du premier intermède est légèrement décalée à cause d'une erreur de SRL concernant le premier présentateur. L'intermède trouvé automatiquement rassemble le jingle de fin du premier programme et le jingle de début du journal.

3 : Une erreur de SRL fait débuter l'intervention du présentateur du journal avec quelques secondes de retard. Cette erreur est liée au fait que le présentateur commence à annoncer les titres sur la fin du jingle.

4 : Le journal est suivi d'une chronique d'une centaine de secondes. Aucun présentateur n'est présent sur cette zone, il s'agit donc d'un intermède.

5 : Le présentateur de la troisième unité de programme commence à parler à la fin du générique musical : le début du troisième programme est très bien retrouvé.

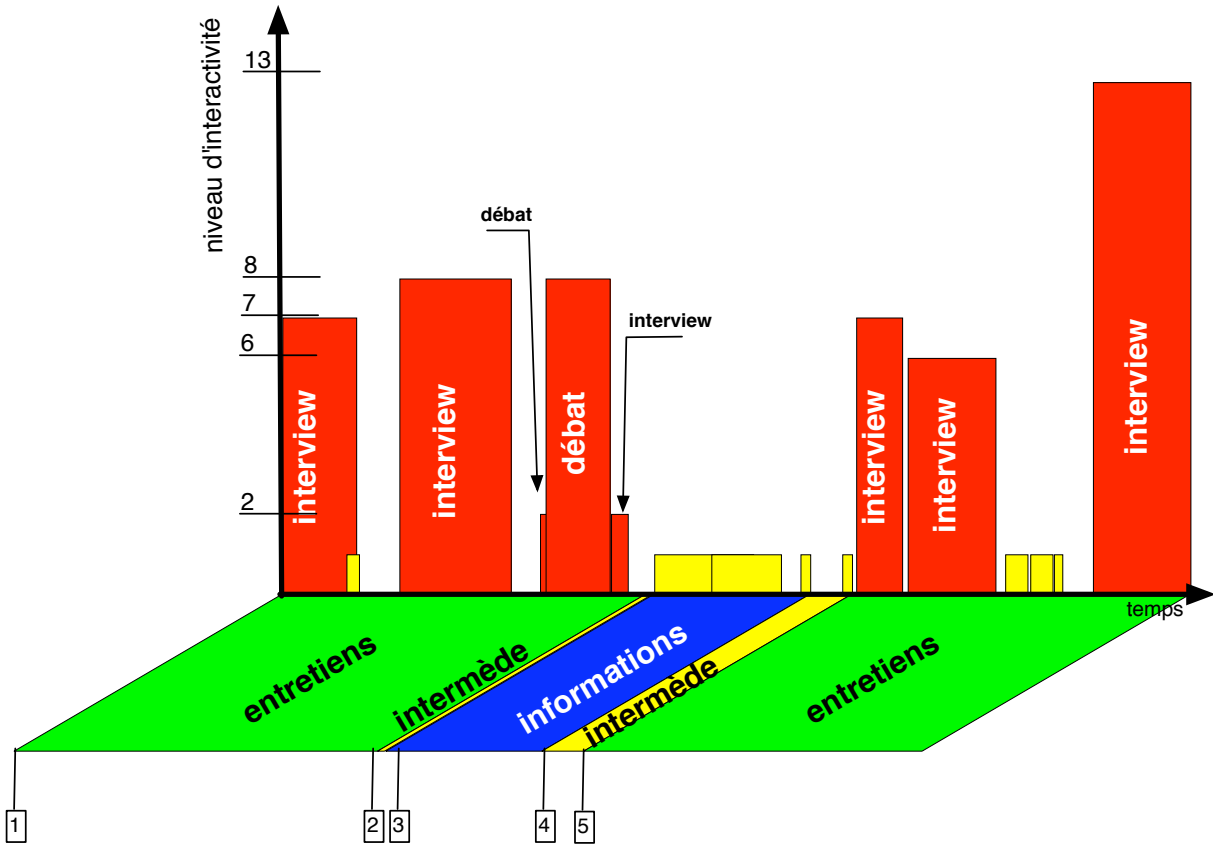


FIGURE 4.15 – Résultats du premier niveau de structuration (frise du bas) et du second niveau de structuration (frise du haut) sur une tranche horaire de RFI contenant plusieurs programmes. Les zones d'interaction apparaissent en jaune quand leur niveau d'interactivité est égal à 1 et apparaissent en rouge si le niveau d'interactivité est supérieur à 1.

Second niveau de structuration. Nous associons la représentation de la figure 4.14 avec la représentation temporelle de zones d'interaction pour obtenir la figure 4.15.

1, 5 : Nous observons que les unités de programmes « entretiens » détectées contiennent plusieurs zones d'interactions de type « interview » et « débat ».

2 : L'unité de programme « informations » ne comporte qu'au plus des zones d'interaction de niveau 1 (représentées en couleur jaune). Ce journal ne contient *a priori* pas de zone conversationnelle présentant une interaction importante.

4.4.3.2 Structuration d'un débat de société

Le Téléphone Sonne est une émission de la station de radio France Inter abordant chaque soir en direct un thème différent en lien avec l'actualité. Les auditeurs de l'émission interviennent par téléphone pour poser des questions à plusieurs invités (journalistes, universitaires, personnalités politiques...), présents en studio ou par téléphone. La figure 4.16 rassemble les résultats obtenus à chaque étape de la structuration de ce programme.

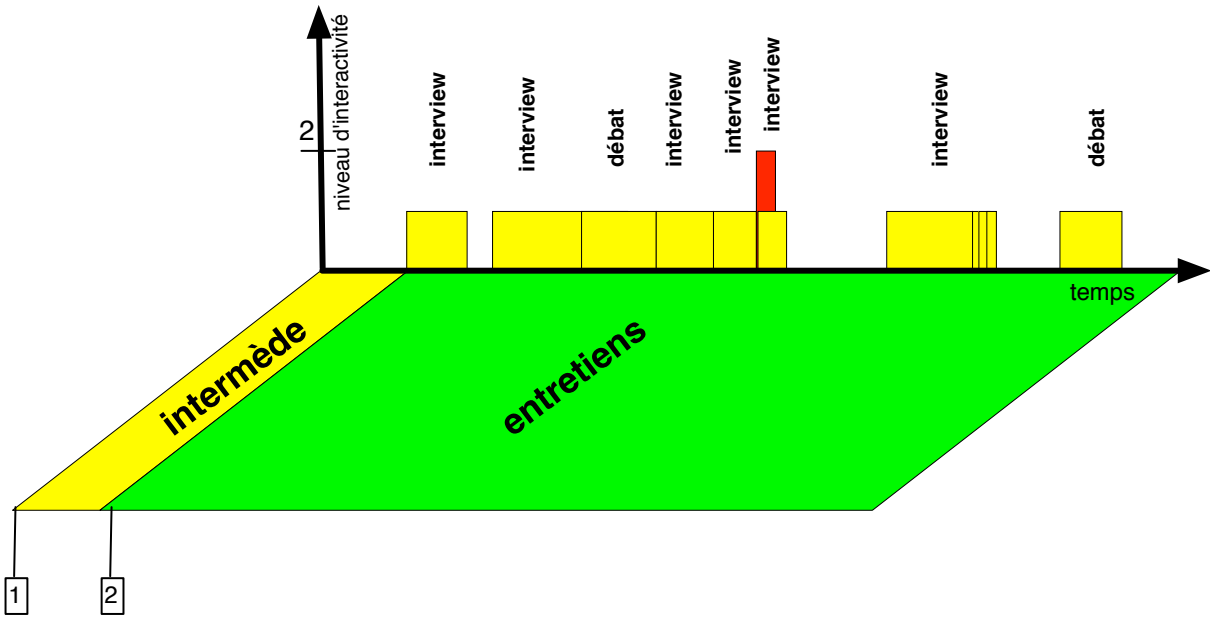


FIGURE 4.16 – Résultats du premier niveau de structuration (frise du bas) et du second niveau de structuration (frise du haut) sur un débat de société *Le Téléphone Sonne* de France Inter. Les zones d'interaction apparaissent en jaune quand leur niveau d'interactivité est égal à 1 et apparaissent en rouge si le niveau d'interactivité est supérieur à 1.

Premier niveau de structuration. Deux unités sont mises en évidence.

1 : Ce premier segment, trouvé comme étant un « intermède », est en réalité l'introduction de l'émission faite par le présentateur pendant un générique musical. Une nouvelle fois, la superposition de parole et de musique a conduit le système à commettre une erreur de SRL. Durant cette séquence le présentateur introduit le thème du débat et les participants.

2 : Le second segment débute après la fin du générique (le premier tour de parole du présentateur détecté) : la structuration est alors parfaite jusqu'à la fin de l'émission.

Second niveau de structuration.

1 : Le premier segment est un monologue : effectivement, aucune zone d'interaction ne correspond à cette unité d'intermède.

2 : Bien que le contenu du programme soit conversationnel, les zones d'interaction découvertes sont presque toutes de niveau d'interaction 1. C'est un résultat assez décevant car la détection d'interaction appliquée à ce même document à partir d'une annotation manuelle révèle de nombreuses zones d'interaction (cf. figure 1.10).

La mauvaise détection des zones s'explique par les erreurs de SRL sur ce contenu très conversationnel. Pourtant, l'erreur globale de segmentation et de regroupement en locuteur est plutôt faible : DER de 8%. Ceci est plutôt le résultat de nombreuses petites erreurs locales, très courtes, ayant pour conséquence de ne pas permettre à la détection des zones d'interaction d'établir des séquences d'alternances très longues entre deux locuteurs. Nous rappelons que nous avions imposé à la détection des z.i. de ne tolérer que des intervalles sans parole de durées strictement inférieures à une seconde. Il est probable que, dans ce contexte de parole très spontanée, cet intervalle soit trop contraignant. En effet, durant ce débat, les invités doivent répondre, sans préparation, aux questions des invités, favorisant l'apparition de disfluences telles que des pauses pleines ou des interjections, des rires, etc. Tout ceci est, sans nul doute, source d'erreurs.

4.4.3.3 Structuration d'une émission de type matinale

Notre approche est appliquée à un enregistrement de la matinale de France Culture. C'est un programme de 120 minutes, animé par un présentateur principal, présent du début à la fin de l'émission. D'autres présentateurs apparaissent au cours de ce programme de telle sorte qu'il est presque possible de comparer la structuration de cette émission à celle d'un document comportant plusieurs programmes.

Premier niveau de structuration. Nous détaillons dans la suite les principales phases de l'émission afin de permettre au lecteur de mettre en relation la structure obtenue et le contenu réel du programme (cf. figure 4.17).

1 : La matinale débute par une introduction du présentateur principal qui n'est pas détectée : comme pour les exemples précédents, c'est une zone de parole superposée à de la musique qui n'a pas été correctement traitée par l'étape de SRL de notre système.

2-4 : Le programme se poursuit avec un bulletin d'information (entre les instants 168 et 611), animé par un second présentateur correctement identifié par le système. Ce bulletin d'information est coupé en deux par une unité « transition » [3], qui correspond à une intervention du présentateur principal dans le journal.

5 : Un premier « entretiens » correspond à une séquence durant laquelle le présentateur principal appelle au téléphone le rédacteur en chef d'un quotidien, puis le segment se termine par un monologue du présentateur qui présente un agenda culturel.

6 : Un second « entretiens » coïncide avec l'apparition d'un nouveau présentateur.

7 : Le présentateur principal reprend la parole et discute avec le présentateur de la séquence précédente : ceci génère la détection d'une unité « transition ».

8 : L'émission se poursuit avec un nouveau journal plus court que le précédent.

9 : Un « entretiens » débute par une chronique lue par un locuteur dont le rôle *journaliste non ponctuel* a été correctement trouvé par le système de reconnaissance des rôles.

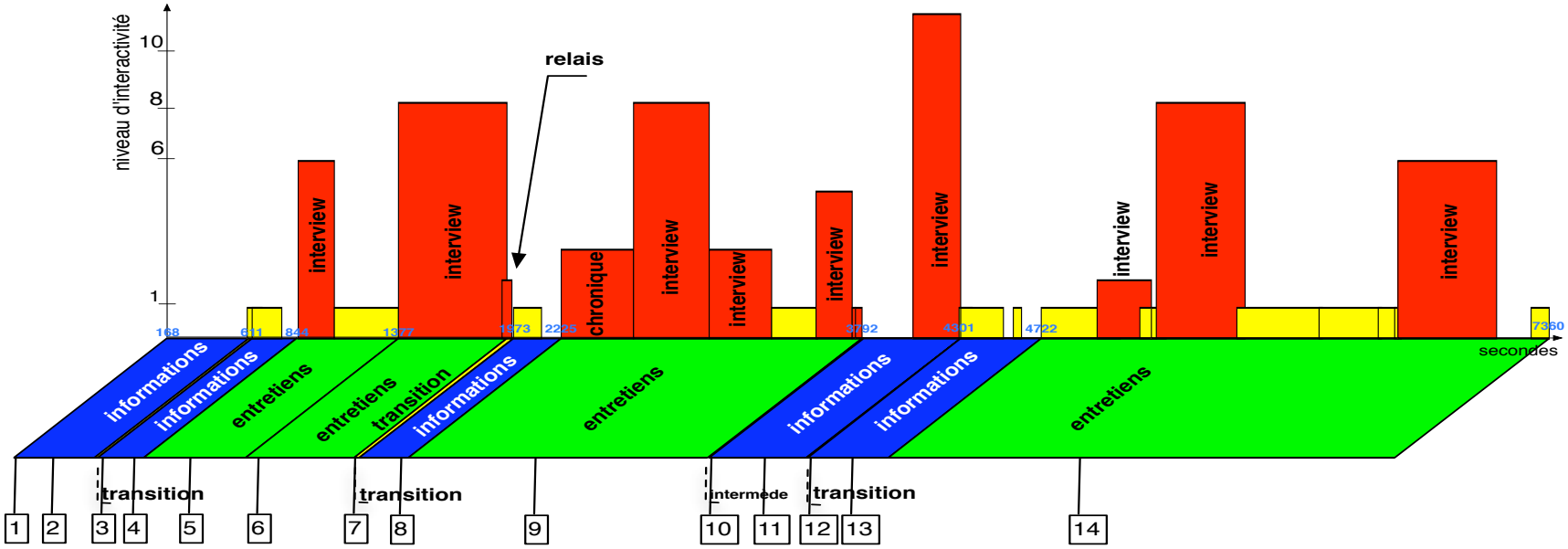


FIGURE 4.17 – Résultats du premier niveau de structuration (frise du bas) et du second niveau de structuration (frise du haut) sur l'émission matinale *Les Matins de France Culture*. Les zones d'interaction apparaissent en jaune quand leur niveau d'interactivité est égal à 1 et apparaissent en rouge sinon le niveau est supérieur à 1.

10 : Un « intermède » correspond au jingle et à l'annonce du journal.

11-13 : Le système détecte tout à fait justement un nouveau programme « informations », coupé en deux parties [11] et [13] par une intervention du présentateur principal [12].

14 : Un dernier « entretiens » est détecté. Celui-ci est réalisé par le présentateur principal et débute par une chronique. Rien ne permet de le voir sur la figure obtenu. Le programme se poursuit par l'interview de l'invité.

Second niveau de structuration.

1-4 : Les premiers segments correspondent à un journal. Aucune zone d'interaction n'est détectée, à juste titre.

5 : Le système a correctement isolé la zone d'interaction de niveau 6, correspondant à l'interview du rédacteur en chef.

6 : Ce programme est un entretien entre un présentateur et un économiste. Le système y détecte une zone d'interaction de niveau 8.

7 : La « transition » entre les deux présentateurs apparaît sous la forme d'une zone d'interaction de niveau supérieur à 1, attribuée à la catégorie « relais ».

9 : La première zone d'interaction de cette unité de programme est une chronique lue par un journaliste et qui débute par une conversation avec le présentateur. Le programme se poursuit avec l'interview d'un biologiste renommé. Deux zones d'interaction adjacentes correspondant à cet échange sont détectées. Cette séparation en deux zones d'interaction est probablement liées à un intervalle trop long (supérieur à une seconde) entre deux tours de parole. Néanmoins, les zones sont correctement attribuées à la catégorie « interviews ».

10-13 : Une zone d'interaction de niveau égal à 11 est présente au milieu du bulletin d'information. Cette zone dure peu de temps, il s'agit fort justement de l'interview très interactionnelle d'un syndicaliste.

14 : Cette zone est la plus longue du document : elle commence par une zone d'interaction de niveau 1 (en jaune) puis débute une première interview (identifiable par une zone d'interaction de niveau 2), suivi d'une zone de niveau 8. Ces z.i. « interviews » sont suivies de deux zones de niveau 1. Ce programme se poursuit avec l'interview d'un nouvel invité : un psychothérapeute. Cet interview correspond à la dernière zone d'interaction de niveau 6. Le programme se termine sur un monologue de conclusion réalisée par le présentateur principal.

Ce dernier exemple est particulièrement intéressant car il met en relief la complémentarité de nos deux niveaux de structuration. Le premier niveau permet de faire émerger la structure du document en zones « entretiens » et « informations ». Dans cet exemple, les noms donnés à ces zones représentent bien leur contenu réel, c'est-à-dire des bulletins d'information dans un cas et principalement des interviews d'invités dans l'autre cas. Le premier niveau de structuration génère des zones « entretiens » assez longues, contenant parfois plusieurs événements. La complémentarité du premier et du second niveau de structuration apparaît clairement quand le résultat du second fait émerger plusieurs séquences conversationnelles cohérentes par leur

contenu. Le parcours non linéaire du document devient possible en passant d'une zone d'interaction à la suivante. En effet, la majorité des zones d'interaction indique une nouvelle séquence de l'émission.

4.5 Conclusion

Dans ce chapitre, nous avons présenté une méthode de structuration automatique des documents audiovisuels, fondée sur la connaissance des rôles des intervenants et leurs interactions orales. La structuration se déroule en deux niveaux de structuration, fondées sur les résultats d'une segmentation en locuteurs enrichie par les rôles des locuteurs. Une macro-segmentation permet d'extraire des documents audiovisuels un ensemble de segments appelés des unités « présentées » et des unités « intermèdes ». Les unités « présentées » sont alors classées en 3 catégories : « entretiens », « informations » et « transition », en fonction des rôles des locuteurs qui parlent le plus sur l'unité. Les « intermèdes » sont considérés comme des unités « inter-programmes » : leur caractérisation s'arrête à ce niveau.

Durant une seconde étape de caractérisation, nous nous focalisons sur les zones d'interaction détectées durant les unités « présentées ». Les zones d'interaction sont caractérisées en quatre catégories : « interview », « chronique », « débat » et « relais ».

Cette méthode de structuration a été intégrée dans un système entièrement automatique, utilisant une méthode de segmentation et de regroupement en locuteurs, notre méthode de détection des zones d'interaction, ainsi que notre système de reconnaissance des rôles. Une évaluation est menée sur les documents de l'ensemble du corpus de test d'EPAC. La reconnaissance atteint un score supérieur à 85% pour le premier niveau de structuration (sur les quatre types d'unités). Le second niveau, quant à lui, obtient un score de reconnaissance supérieur à 67% (sur les quatre catégories).

Nous avons également illustré, à travers plusieurs exemples, la complémentarité des deux niveaux de structuration. Le premier niveau permet une macro-segmentation cohérente, grâce au pouvoir d'ancrage du présentateur. Il s'agit d'un niveau de granularité permettant le parcours du contenu à l'échelle du flux audiovisuel. Le second niveau de structuration présente une granularité plus fine : micro-segmentation. Il permet de mettre en évidence, à l'intérieur des unités « présentées », des zones caractéristiques correspondant à des séquences conversationnelles.

Conclusion et perspectives

1 Bilan de nos travaux

La contribution de cette thèse s'inscrit dans le domaine de l'indexation et de la structuration des documents audiovisuels. Nous nous sommes intéressés plus particulièrement à l'exploitation d'informations accessibles à travers les rôles des intervenants et les interactions orales. Ce travail de thèse a été financé par l'ANR dans le cadre du projet EPAC « Exploration de masses de documents audio pour l'extraction et le traitement de la PArole Conversationnelle ».

Les débuts de nos travaux donnent suite à l'étude menée par Ibrahim autour de la caractérisation des relations temporelles observées au cours d'une conversation, entre les tours de parole de locuteurs [Ibrahim 07]. Les résultats établis par cette étude antérieure nous ont amené à proposer **une méthode de détection des zones d'interaction**, fondée sur la recherche des séquences d'alternance des tours de parole de deux locuteurs. La détection de telles séquences est réalisée à partir du résultat d'une segmentation et d'un regroupement en locuteurs (SRL). Nous utilisons la SRL développée par El Khoury [El Khoury 10] en pré-traitement. Notre méthode ne fait aucune hypothèse sur le contenu linguistique des conversations détectées, nous nommons donc une telle séquence détectée **une zone d'interaction orale** et nous lui associons **un niveau d'interactivité**. Il s'agit d'une mesure simple indiquant le potentiel conversationnel de la zone d'interaction par le calcul du nombre d'alternances de tours de parole correspondant.

Nous réalisons par la suite **une caractérisation des locuteurs qui apparaissent dans les zones d'interaction.** Durant ce travail, nous avons mis en relief les particularités des locuteurs ponctuels qui sont des locuteurs n'apparaissant que sur un seul segment ou tour de parole. Nous proposons, pour caractériser les locuteurs, plusieurs descripteurs temporels globaux et locaux :

- les **descripteurs locaux**, à travers *l'activité locale* et *le vecteur de répartition de l'activité*, mesurent la répartition des interventions de parole d'un locuteur tout au long d'un document. Nous avons évoqué l'intérêt de ces descripteurs dans une perspective de travail sur la caractérisation des documents en fonction des vecteurs de répartition observés.
- les **descripteurs globaux** caractérisent les locuteurs grâce à *l'activité globale* et *l'étendue*.

Dans ce travail préliminaire, les descripteurs de l'activité globale et de l'étendue nous permettent de proposer une première typologie des interventions des locuteurs. Nous déclinons cinq catégories d'interventions, définies en fonction des valeurs des deux descripteurs globaux précités. Une observation empirique des locuteurs rassemblés dans ces catégories semblent indi-

quer *a priori* un lien entre les catégories proposées et les rôles réels des intervenants. Ce résultat est à l'origine de nos travaux autour de la reconnaissance automatique des rôles des intervenants.

Une autre contribution concerne **l'étude de paramètres pertinents pour la reconnaissance des rôles.** Les travaux de la littérature de ce domaine s'accordent sur la reconnaissance des rôles : présentateur, journaliste et autre (ou invité). Nous utilisons ces trois rôles également, et nous proposons de les faire évoluer **vers cinq rôles**, en intégrant la distinction évoquée auparavant entre les locuteurs ponctuels et non ponctuels. Les rôles que nous étudions sont finalement : *présentateur*, *journaliste non ponctuel*, *journaliste ponctuel*, *autre non ponctuel*, *autre ponctuel*.

Afin de capter les caractéristiques de ces rôles, nous proposons de **représenter chaque intervenant à travers 36 paramètres de « bas-niveau »,** accessibles à partir du résultat d'une SRL. Cet ensemble de paramètres rassemble :

- 14 paramètres temporels, calculés à partir des segments de parole issus de la SRL,
- 10 paramètres acoustiques, calculés directement sur le signal audio,
- 12 paramètres prosodiques, calculés à partir d'une estimation de la fréquence fondamentale et d'une segmentation vocalique.

La pertinence de cet ensemble de paramètres est évaluée à travers l'utilisation d'une **méthode non supervisée sur les vecteurs de 36 paramètres** d'un ensemble de locuteurs de rôles connus. Les regroupements obtenus sur le corpus de test du projet EPAC avec la méthode des K-means (K=20) rapportent une pureté en rôles moyenne de 80% pour cinq rôles. Ce résultat très bon est corroboré par la manière dont se répartissent les clusters parmi les cinq rôles considérés. La pertinence des 36 paramètres ayant été illustrée, nous les intégrons à notre système de reconnaissance automatique de rôle.

La suite de notre travail se concentre autour de **l'étude d'un système de reconnaissance automatique des rôles des locuteurs.** Fondé sur la structure classique d'un système de reconnaissance des formes, il se compose :

- d'une phase de pré-traitement des données (la SRL dans notre cas),
- d'une extraction de 36 paramètres de « bas-niveau »,
- d'une éventuelle réduction de dimension des données. Nous considérons tour à tour l'analyse en composantes principales (ACP), l'analyse factorielle discriminante (AFD) et une méthode de sélection de paramètres par recherche séquentielle par élimination (RSE).
- d'une classification supervisée : les GMM, les k-ppv et les SVM à noyaux linéaires, rbf, sigmoïdal ou polynomial sont étudiés.

Nous avons évalué plusieurs variantes de ce système à travers de nombreuses expériences. Celles-ci sont réalisées à l'aide de deux ensembles de documents :

- le corpus ESTER2 qui contient en majorité des bulletins d'information,
- le corpus EPAC qui comporte une proportion importante de parole conversationnelle (débats, interviews et magazines).

Les premières expériences réalisées concernent la reconnaissance des trois rôles classiques de la littérature : *présentateur*, *journaliste* et *autre*. Les performances sont exprimées en taux de reconnaissance correcte (TRC) accompagnés de leur intervalle de confiance à 95%. L'un des premiers systèmes retenu pour ses performances intègre un SVM à noyau rbf. Il permet de reconnaître le rôle pour $\mathbf{79,3\% \pm 5,6}$ des locuteurs du corpus ESTER2. Ce score est comparable aux performances rapportées dans la littérature : entre 80% et 85% de rôles bien reconnus.

Nous proposons une seconde série d'expériences, intégrant les cinq rôles issus de la distinction entre les locuteurs ponctuels et non ponctuels. Le système à cinq rôles le plus abouti est fondé sur une architecture hiérarchique. Il se compose d'un classifieur SVM linéaire et d'une étape de sélection de paramètres par éliminations successives. Le taux de reconnaissance le plus élevé est obtenu sur le corpus d'ESTER2 : $\mathbf{81,3\% \pm 5,4}$. Appliqué aux données du corpus EPAC, ce même système rapporte un taux de $\mathbf{83,2\% \pm 7,1}$.

L'architecture hiérarchique permet d'analyser les paramètres les plus discriminants à chaque étape de la classification. Sur le corpus ESTER2, il s'agit de :

- la variance de la puissance du signal sur les zones attribuées au locuteur,
- la valeur maximale du pitch,
- le nombre de silences,
- la durée moyenne des silences.

Sur le corpus EPAC, les paramètres retenus sont :

- la puissance moyenne du signal sur les zones de non parole,
- la valeur minimale de la puissance du signal sur les zones de parole,
- le nombre de silences par unité de temps,
- la variance de la durée des silences.

Ce résultat illustre d'une part l'importance des descripteurs prosodiques dans la reconnaissance des rôles, sans pour autant qu'ils soient identiques pour chaque ensemble de données. Ce dernier point relève de la spécificité des locuteurs contenus dans chacun des ensembles de données.

Finalement, le meilleur score de classification est obtenu sur le corpus EPAC à l'aide du système hiérarchique intégrant une ACP et un classifieur SVM linéaire. Dans ce cas, $\mathbf{92\% \pm 5,3}$ des rôles sont correctement attribués aux locuteurs. Cet excellent résultat nous permet d'utiliser ce système de reconnaissance de rôles, dans notre système complet de structuration des documents audiovisuels. Notons que la reconnaissance du présentateur est très robuste : celle-ci est quasi parfaite, quel que soit le corpus (ESTER2 ou EPAC).

Notre dernière contribution est **une méthode de structuration automatique des documents audiovisuels**. Cette méthode est fondée sur la connaissance *a priori* des rôles des intervenants et leurs interactions orales. La structuration se déroule en deux étapes, fondées sur les résultats d'une SRL enrichie par les rôles des locuteurs et leurs interactions.

Au premier niveau de structuration, une macro-segmentation permet d'extraire des documents audiovisuels un ensemble de segments appelés unités « présentées » et unités « intermèdes ». Les unités « présentées » sont alors classées en 3 catégories : « entretiens », « informa-

tions » et « transition », en fonction des rôles des locuteurs qui y parlent le plus. Les « intermèdes » sont considérés comme des unités « inter-programmes » : leur caractérisation s'arrête à ce niveau.

Durant une seconde étape de caractérisation, nous nous focalisons sur les zones d'interaction détectées durant les unités « présentées ». Les zones d'interaction sont classées en quatre catégories : « interview », « chronique », « débat » et « relais ».

Cette méthode de structuration automatique des documents audiovisuels a été intégrée dans un système entièrement automatique (cf. figure 1), dont voici les différentes étapes :

(1) la SRL de El Khoury [El Khoury 10] : cette première étape fournit un ensemble de segments temporels étiquetés avec des identifiants de locuteurs loc_j,

(2) la méthode de détection des zones d'interaction orale entre intervenants. Celles-ci sont détectées à partir de la recherche de séquences d'alternances dans les segments issus de l'étape de SRL. Le résultat de ce traitement est un ensemble de zones temporelles, associées à leur niveau d'interactivité respectif.

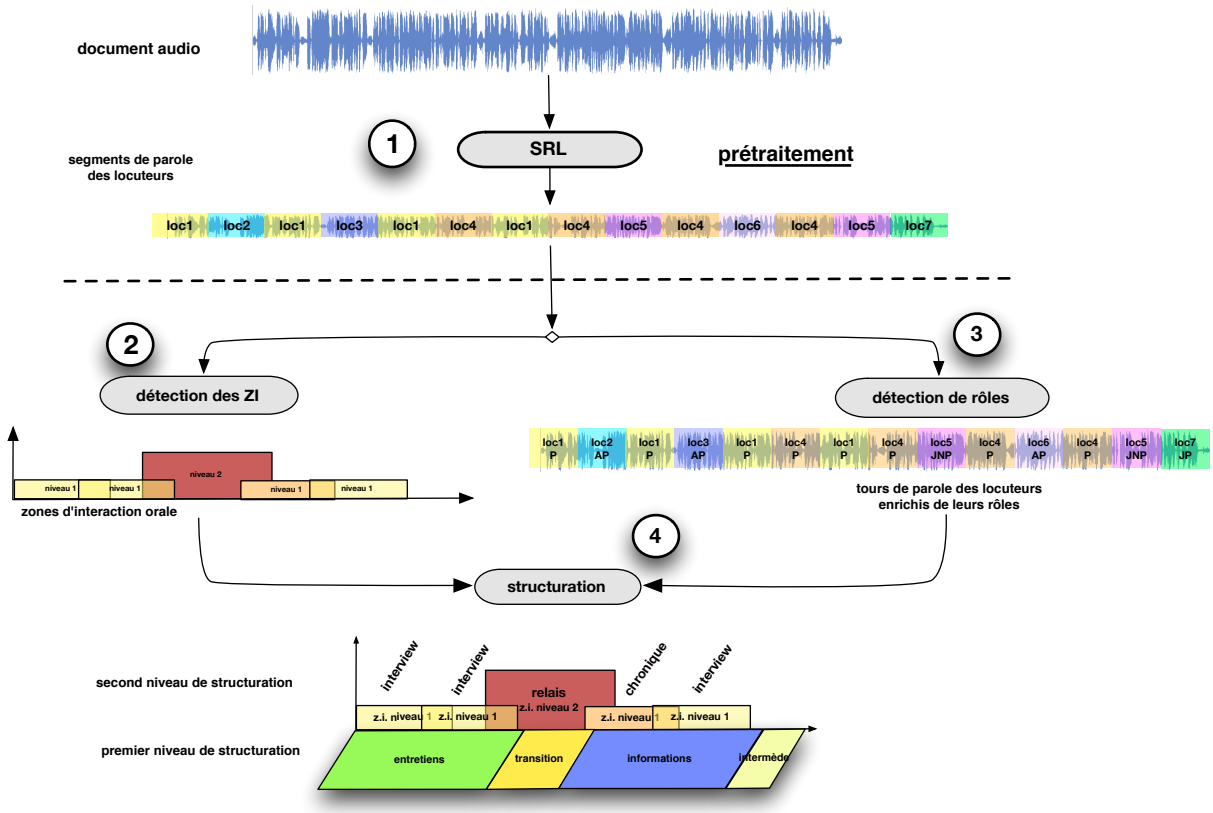


FIGURE 1 – Les quatre étapes de traitement du système complet de structuration automatique des documents sonores.

(3) le système hiérarchique de reconnaissance automatique des rôles est composé d'une réduction de la dimension (par ACP) et d'un classifieur SVM linéaire. Ce traitement fournit une segmentation en locuteurs, enrichie du rôle de chacun.
(4) le système de structuration à deux niveaux, intégrant les résultats des étapes précédentes.

Une évaluation est menée sur les documents de l'ensemble du corpus de test d'EPAC. La reconnaissance atteint un score supérieur à 85% pour le premier niveau de structuration (sur les quatre types d'unités). Le second niveau, quant à lui, obtient un score de reconnaissance supérieur à 67% (sur les quatre catégories).

Nous avons également illustré, à travers plusieurs exemples, la complémentarité des deux niveaux de structuration. **Le premier niveau permet une macro-segmentation** cohérente, grâce au pouvoir d'ancrage du présentateur. Il s'agit d'un niveau de granularité permettant le parcours du contenu à l'échelle du flux audiovisuel. **Le second niveau de structuration présente une granularité plus fine : micro-segmentation**. Il permet de mettre en évidence, à l'intérieur des unités « présentées », des zones caractéristiques du contenu, correspondant à des séquences conversationnelles.

2 Perspectives

Notre système complet de structuration présente un potentiel d'évolution important. Chaque partie de son architecture, que nous nommons sous-système, peut être le support de recherches spécifiques et chaque amélioration locale d'un des sous-systèmes pourrait être bénéfique au système complet. Dans cette section, consacrée à la présentation de nos perspectives de recherche, nous commençons par évoquer quelques propositions des plus immédiates puis nous élargissons le champ de nos perspectives de travail.

2.1 Améliorations à court terme des sous-systèmes

Les perspectives de recherche à court terme consistent dans quelques propositions en vue d'augmenter les performances de chacun de nos quatre sous-systèmes : SRL, détection des zones d'interaction, détection de rôles et structuration.

La segmentation et le regroupement en locuteurs : nous constatons une erreur fréquente de regroupement sur les zones de parole et de musique superposées. D'après l'auteur de l'algorithme de SRL [El Khoury 10], et comme nous avons également pu le constater dans notre étude, de nombreuses erreurs de regroupement sont liées à la présence de zones de parole superposées à de la musique. Des améliorations sont cependant possibles en tenant compte d'indicateurs internes au module de SRL. En effet, ceux-ci combinés avec la connaissance sur le rôle des locuteurs, voire la structure en macro-segments, pourraient affiner les résultats du regroupement.

La détection des zones d'interaction : lorsque nous avons présenté la méthode de détection des zones d'interaction, nous avons indiqué qu'une valeur de seuil limitait l'intervalle temporel maximal devant séparer les tours de parole de locuteurs successifs, présents dans une

zone d'interaction. Dans le but d'assurer la cohérence des alternances de tours de parole, nous avons utilisé une valeur d'intervalle assez faible, fixée par défaut à une seconde. Cette valeur s'est révélée être adapté dans un grand nombre de situations, mais dans un contexte conversationnel spontané cet intervalle temporel séparant les tours de parole des intervenants peut être amené à augmenter au-delà d'une seconde. Nous souhaitons étudier, à partir d'une segmentation de référence, puis d'une segmentation automatique, l'influence de ce paramètre sur le nombre de zones d'interaction détectées.

Nous avons également commencé à travailler sur une méthode de détection de zones d'interaction impliquant plus de deux personnes, fondée sur les mêmes définitions fondamentales (unités d'interaction) que la méthode actuelle. Cette méthode se fonde sur l'hypothèse qu'une zone d'interaction entre N intervenants contient au moins une unité d'interaction pour tous les couples de locuteurs. Ainsi une zone d'interaction commune à trois locuteurs (loc_1, loc_2 et loc_3), telle que nous la définissons est la plus longue séquence de tours de parole de ces locuteurs. Cette séquence devant contenir au moins une unité d'interaction des couples de locuteurs $\{loc_1$ - $loc_2\}$, $\{loc_1$ - $loc_3\}$, et $\{loc_2$ - $loc_3\}$.

La reconnaissance automatique des rôles : une particularité importante de notre méthode de reconnaissance réside dans son indépendance vis-à-vis de la structure du document. Dans cette méthode, aucune information *a priori* sur le type de document ou le type de programme n'est utilisée. Notre proposition repose sur le fait qu'un intervenant est représenté par un vecteur de 36 paramètres « bas-niveau », indépendamment de toute autre information. Nous observons que certaines erreurs de reconnaissance des rôles pourrait être facilement détectées, notamment lorsque dans un même document plusieurs locuteurs se sont vus attribué le rôle de présentateur. Plusieurs situations sont alors possibles :

- de présentateurs appartenant à des émissions successives ou temporellement disjointes,
- de présentateurs présentant conjointement une même émission,
- de présentateurs dont les niveaux hiérarchique sont différents comme nous avons pu l'observer dans les émissions de type matinale qui comptent un présentateur principal rejoint par des présentateurs secondaires,
- d'une erreur de reconnaissance du rôle.

Une perspective de nos travaux à court terme concerne l'étude de chacune de ces situations en vue de les détecter et de vérifier en amont de l'étape de structuration s'il s'agit d'une erreur de reconnaissance du rôle ou non.

Dans l'objectif d'améliorer les performances de la reconnaissance automatique des rôles, nous souhaitons dans un futur proche, évaluer l'apport d'une combinaison de classifieurs.

La possibilité de faire émerger de l'ensemble des locuteurs de rôles *autre* plusieurs sous-catégories de rôles est également un objectif à court terme. En effet, cette catégorie regroupe une grande diversité de locuteurs, allant d'un invité expert au simple auditeur posant une question. La connaissance du statut de ces locuteurs permettrait d'affiner leurs rôles et ainsi de cerner un peu mieux différentes catégories d'intervention. Cela nécessite d'analyser le discours et notamment ce qui est dit en introduction, lors de l'ouverture des zones d'interactivité. Ces zones permettraient de localiser les positions de l'émission susceptibles de contenir les informations sur le nom et le statut de l'invité par exemple ou la spécialité du journaliste.

La structuration audiovisuelle : les unités « entretiens », « informations », « transition » et « intermèdes », ont été définies de manière à ce que ce premier niveau de structuration reste applicable à plusieurs type de documents et de programmes. Une perspective de recherche consiste à étudier les unités rassemblées sous une même catégorie dans le but de faire émerger une déclinaison plus fine des contenus, et ainsi d'introduire un niveau de structuration supplémentaire au résultat actuel. L'impact des modifications apportées par le premier niveau de structuration sur le second niveau sera étudiée.

2.2 Enrichissement de la structuration fondée sur des événements audio

Notre système de structuration repose uniquement sur la connaissance d'informations extraites des tours de parole des intervenants rendus disponibles par une SRL. À l'aide de la connaissance des rôles des intervenants et des zones d'interaction nous sommes allés assez loin dans la caractérisation du contenu. En accord avec notre problématique de départ, nous ne souhaitons pas introduire d'informations linguistiques pour le moment. Une possibilité afin de progresser dans la catégorisation des unités de structuration extraites par notre méthode, peut être d'introduire des informations liées aux événements sonores fréquents dans un grand nombre de programmes, tels que la musique, les jingles, des rires ou des applaudissements.

L'intégration d'informations supplémentaires concernant la présence et la localisation d'événements sonores variés est une perspective intéressante qui devrait nous permettre d'évoluer vers un nouveau niveau de structuration proche du genre des programmes dans le meilleur cas, à l'instar du résultat de structuration présenté sur la figure 2. Sa mise en parallèle avec les résultats fournis par une méthode de segmentation en musique, rires et applaudissement peut faire évoluer la compréhension que nous avons du contenu de l'unité « présentée ».

Bien évidemment, nous pourrions exploiter également les intermèdes (publicités, promotions, sponsors...).

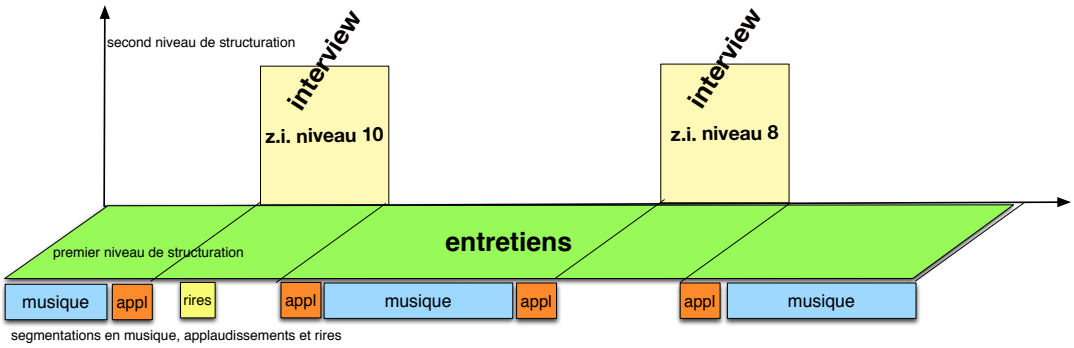


FIGURE 2 – Un exemple d'unité « présentée » de type « entretiens » à laquelle s'ajoute le résultat que pourrait fournir un algorithme de segmentation en musique, rires et applaudissements.

2.3 Exploitation d'informations extraites de la vidéo pour la structuration

Il est assez courant dans les débats de société ou politique diffusés à la télévision, de trouver des séquences de parole superposée. Au contraire des programmes radiophoniques dans lesquels il

FIGURE 3 – Intervention audiovisuelle d'un présentateur dans le contexte d'un plateau.

FIGURE 4 – Intervention d'un journaliste en voix-off dans le contexte d'un reportage.

est indispensable que les intervenants respectent la parole des autres pour garantir l'intelligibilité du programme, les intervenants à la télévision bénéficient d'une liberté d'interagir plus grande : l'information visuelle permet dans ce cas de compenser les limites de l'audio.

Une perspective de notre travail concernera directement l'exploitation de l'information disponible dans le flux audiovisuel afin de « soutenir » notre système fondé sur une approche purement audio. Les points sensibles sur le contenu vidéo, que nous devrons rendre plus robustes sont d'une part la SRL, et d'autre part la reconnaissance automatique des rôles. Nous pensons plus particulièrement intégrer les éléments suivants :

- au niveau de la SRL : le travail réalisé par El Khoury en 2010 concerne la segmentation et le regroupement en intervenants dans la vidéo. Ceux-ci montrent que la fusion du résultat d'une SRL, telle que celle que nous utilisons, avec un détecteur de visages parlants, permet d'améliorer grandement les résultats. Nous étudierons plus particulièrement les méthodes pouvant permettre d'améliorer cette étape de notre système.
- à l'étape de reconnaissance des rôles : nous souhaitons introduire des informations liées au contexte dans lequel apparaissent certains rôles comme nous l'illustrons à travers deux exemples. La première image (cf. figure 3) représente une séquence de plateau sur laquelle intervient le présentateur de l'émission. La seconde image (cf. figure 4) illustre une séquence de reportage durant laquelle un journaliste parle en voix-off.
  Nous voyons de quelle manière des rôles différents peuvent correspondre également à des contextes différents. Dans le premier cas, il s'agit d'un présentateur, apparaissant uniquement sur des scènes de plateau. Les segments de parole de ce locuteur correspondent avec les séquences d'images sur lesquelles il apparaît. Dans le second cas, il s'agit d'un journaliste, ses interventions locales n'apparaissent que durant des reportages et cet intervenant n'apparaît pas à l'écran durant ses interventions.

Il sera intéressant d'étudier le lien entre ces événements en perspective d'une adaptation à la vidéo, de notre système de reconnaissance des rôles.

2.4 Pistes de recherche autour de la caractérisation locale des intervenants

Dans le chapitre 2, nous avons proposé deux descripteurs permettant de caractériser localement les interventions des locuteurs : l'activité locale et le vecteur de répartition locale. Une perspective de notre travail consiste à faire évoluer cette proposition afin d'étudier localement les descripteurs temporels, acoustiques et prosodiques.

Une première étape de ce traitement pourra consister, comme nous l'illustrons sur la figure 5, à découper le document en plusieurs sections de même durée (trois sections pour l'exemple présenté). Puis pour chaque locuteur du document, évaluer à l'intérieur de ces fenêtres les paramètres acoustiques, prosodiques et temporels. Nous serons en mesure d'observer l'évolution temporelle de ces paramètres et d'étudier leur lien avec les rôles des intervenants, ou de mieux caractériser un rôle en fonction de l'évolution temporelle de ces paramètres. Ce dernier point est d'ailleurs à l'étude actuellement à travers un stage de master recherche.

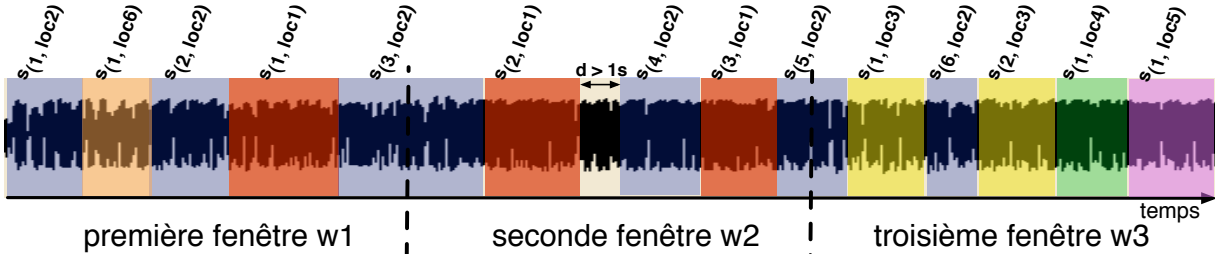


FIGURE 5 – Pré-découpage d'un document en zones d'analyse sur lesquelles les paramètres temporels, acoustiques et prosodiques peuvent être évalués localement.

2.5 Investigation sur une collection de documents fondée sur la théorie des graphes

Nous fermerons la section consacrée à nos perspectives de recherche par la proposition suivante. Elle concerne l'utilisation des rôles des intervenants et des interactions orales qui les lient, dans le but de caractériser les documents dans un premier temps, puis les programmes contenus dans ces documents.

Nous représentons deux exemples sous forme de graphes, rapporté sur la figure 6, pour le cas d'un document contenant deux émissions (magazine et journal), et sur la figure 7, pour le cas d'un journal d'information. Sur ces graphes, un nœud représente un intervenant, pour lequel nous avons indiqué le rôle. La surface des nœuds est proportionnelle au temps de parole de l'intervenant dans le document. Les arcs relient deux intervenants : ils illustrent l'existence d'une interaction entre ces deux locuteurs. Ces arcs sont pondérés par le nombre d'alternances de tours de parole existant entre ces deux locuteurs.

Un objectif dans un premier temps pourra être de générer des collections en recherchant automatiquement dans une masse de documents, des documents présentant des caractéristiques similaires en termes de nombre de densité ou de centralité du présentateur par exemple.

Il sera dans ce cas nécessaire de développer une étude complète de manière à établir une méthode de calcul de la similarité entre deux documents.

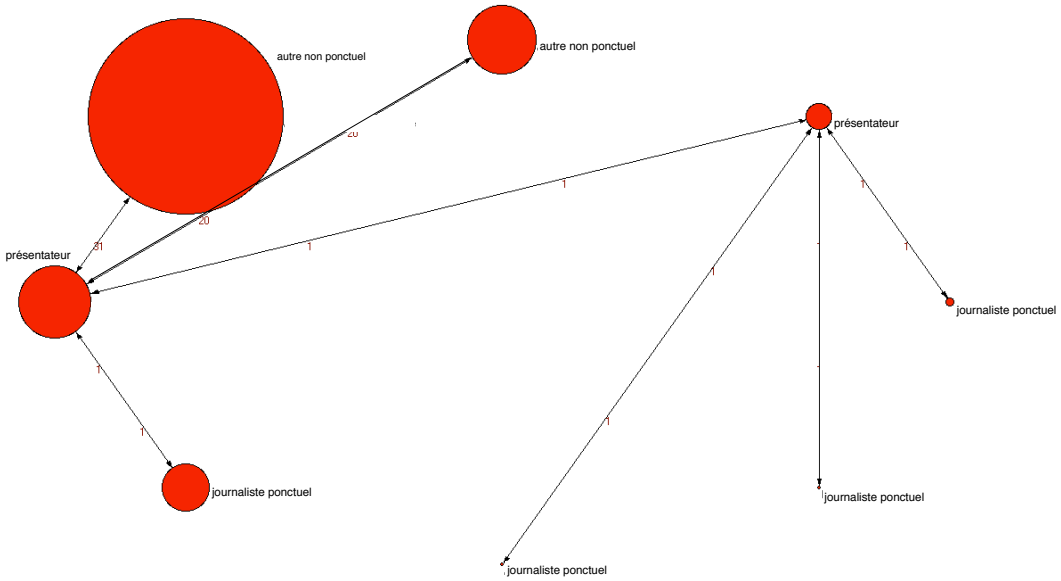


FIGURE 6 – Graphe représentant les intervenants et leurs interactions dans un document contenant deux émissions successives : un magazine et un journal.

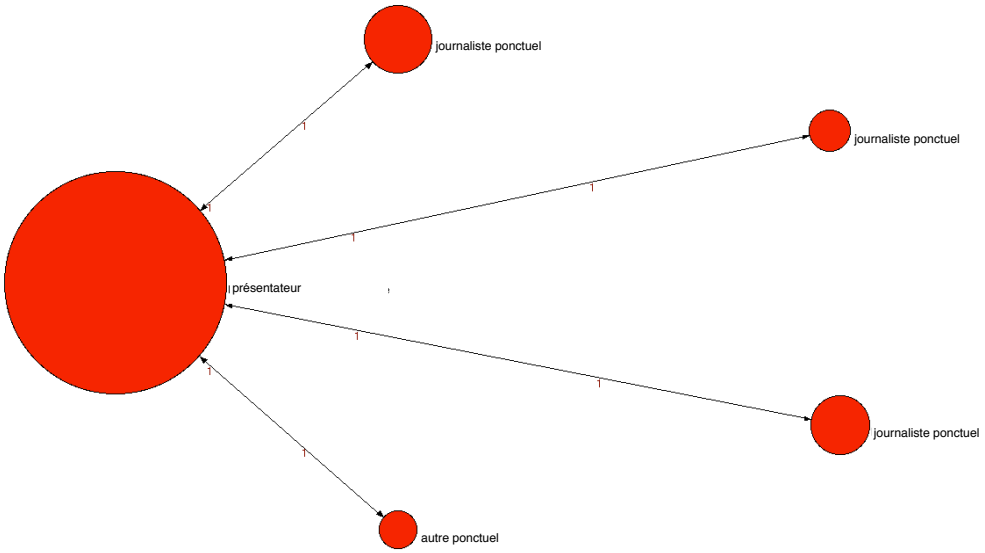


FIGURE 7 – Graphe représentant les intervenants et leurs interactions dans un document contenant un journal.

Annexe A

Système de reconnaissance des rôles

Cette annexe présente les tables regroupant les scores des différentes expériences menées au cours de nos travaux sur la reconnaissance de rôles. Les performances sont exprimées en termes de Taux de Reconnaissance Correcte (TRC) avec l'intervalle de confiance à 95% correspondant, ainsi qu'en proportion de durée bien classée τ. Les performances les plus élevées en termes de taux de reconnaissance correcte sont présentées en gras dans chaque table. Les matrices de confusion détaillent les résultats de la variante du système donnant les scores les plus élévés.

A.1 Étude de l'influence des erreurs de SRL

Dans cette section, sont rassemblées les expériences réalisées à partir du système de reconnaissance à 3 rôles, (1) appliqué à l'intégralité des 36 paramètres, ou avec réduction de dimension (2) ACP et (3) AFD.

A.1.1 SRL manuelle

La table A.1 présente les résultats obtenus à partir de la segmentation en locuteurs manuelle, disponible sur le corpus ESTER2.

TABLE A.1 – Performances du système à 3 rôles : évaluation sur les segmentations manuelles du corpus ESTER-tst, (1) sans réduction de dimension, ou après (2) ACP, ou (3) AFD.

	(1) espace intial 36 paramètres		(2) après ACP 20 dimensions		(3) après AFD 2 dimensions	
	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$
Modèle de Gaussienne	$80,2 \pm 5,3$	83,5	$74,6 \pm 5,8$	67,9	$\mathbf{81,1 \pm 5,2}$	80,6
k-ppv	$71,0 \pm 6,1$	67,9	$71,0 \pm 6,1$	68,5	$80,2 \pm 5,3$	78,7
SVM rbf	$85,7 \pm 4,7$	84,6	$\mathbf{87,6 \pm 4,4}$	83,8	$80,2 \pm 5,3$	73,7
SVM polynomial	$\mathbf{86,6 \pm 4,5}$	84,6	$79,7 \pm 5,4$	77,6	$77,0 \pm 5,6$	80,8
SVM sigmoïdal	$83,9 \pm 4,9$	**88,0**	$80,2 \pm 5,3$	85,8	$80,6 \pm 5,3$	82,6
SVM linéaire	$83,9 \pm 4,9$	**88,0**	$82,9 \pm 5,0$	**86,6**	$80,6 \pm 5,3$	**83,0**

A.1.2 SRL automatique

La table A.2 présente les résultats obtenus à partir de la segmentation en locuteurs automatique sur le corpus ESTER2.

TABLE A.2 – Performances du système à 3 rôles : évaluation sur les segmentations automatiques du corpus ESTER-tst, (1) sans réduction de dimension, ou après (2) ACP, ou (3) AFD.

	(1) espace intial 36 paramètres		(2) après ACP 20 dimensions		(3) après AFD 2 dimensions	
	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$
Modèle de Gaussienne	$72,4 \pm 6,2$	77,6	$72,4 \pm 6,2$	65,6	$78,8 \pm 5,6$	83,1
k-ppv	$63,1 \pm 6,7$	63,0	$64,5 \pm 6,6$	63,9	$74,4 \pm 6,0$	77,5
SVM rbf	$\mathbf{79,3 \pm 5,6}$	**82,8**	$\mathbf{77,8 \pm 5,7}$	77,8	$\mathbf{79,8 \pm 5,5}$	**85,7**
SVM polynomial	$79,3 \pm 5,6$	81,1	$71,4 \pm 6,2$	71,0	$73,4 \pm 6,1$	79,5
SVM sigmoïdal	$75,9 \pm 5,9$	80,7	$74,9 \pm 6,0$	81,3	$78,8 \pm 5,6$	83,5
SVM linéaire	$75,9 \pm 5,9$	80,7	$76,4 \pm 5,9$	**82,4**	$79,3 \pm 5,6$	84,1

À noter : les paramètres retenus pas AFD, maximisant la séparation des classes sur l'ensemble d'apprentissage, sont : $\overline{I}_{seg}$, $\overline{L}_{seg}$, $min(L_{seg})$, $min(P_{env})$, $max(P_{loc})$ et $\overline{P}$ (voir table 2.2 pour la signification de ces paramètres).

A.2 Étude de l'influence des jeux de paramètres temporels, acoustiques et prosodiques

Dans cette section sont rassemblées les expériences réalisées à partir du système de reconnaissance à trois rôles testé avec différents sous-ensembles de paramètres, appliqué à (1) l'intégralité des 36 paramètres, ou avec réduction de dimension (2) ACP et (3) AFD. Les données utilisées sont celles du corpus ESTER2 et SRL automatique.

A.2.1 Paramètres acoustiques

La table A.3 correspondant aux performances de reconnaissance atteintes à partir des paramètres acoustiques seuls. La table A.4 présente la matrice de confusion pour le meilleur système à trois rôles fondé sur l'utilisation des paramètres acoustiques seuls et l'utilisation du corpus ESTER2 et de la SRL automatique.

Notons que ce type de paramètres pris isolément ne permet pas de démarquer franchement une catégorie par rapport à une autre. En effet, nous constatons une grande confusion des locuteurs de type *présentateur* avec la catégorie *journaliste*. Idem entre *journaliste* et *autre*.

A.2.2 Paramètres temporels

La table A.5 rassemble les performances atteintes en reconnaissance par le système à trois rôles avec l'utilisation des paramètres temporels seuls. La table A.6 présente la matrice de

confusion pour le meilleur système à trois rôles fondée sur l'utilisation des paramètres temporels seuls et évalué sur ESTER-tst.

Notons que les paramètres temporels ont un fort impact sur la différenciation des locuteurs *présentateur* par rapport aux deux autres catégories. Nous constatons également un changement au niveau de la catégorie *journaliste* par rapport à *autre*. Par contre, il y a quasiment autant de confusion entre *autre* et *journaliste* qu'avec les paramètres acoustiques.

TABLE A.3 – Performances du système à 3 rôles utilisant les paramètres acoustiques seuls : évaluation sur ESTER-tst, (1) sans réduction de dimension, ou après (2) ACP, ou (3) AFD.

	(1) espace intial		(2) après ACP		(3) après AFD	
	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$
Modèle de Gaussienne	$35,0 \pm 6,6$	43,3	$48,3 \pm 6,9$	$\mathbf{54,0}$	$\mathbf{57,6 \pm 6,8}$	$\mathbf{56,1}$
k-ppv	$52,7 \pm 6,9$	$\mathbf{44,2}$	$49,3 \pm 6,9$	36,9	$52,7 \pm 6,9$	38,3
SVM rbf	$50,7 \pm 6,9$	43,9	$48,8 \pm 6,9$	38,4	$53,7 \pm 6,9$	40,0
SVM polynomial	$42,8 \pm 7,4$	41,3	$44,3 \pm 6,8$	29,7	$46,3 \pm 6,9$	38,0
SVM sigmoïdal	$53,7 \pm 6,9$	39,6	$\mathbf{56,2 \pm 6,8}$	40,9	$53,7 \pm 6,9$	39,0
SVM linéaire	$\mathbf{54,7 \pm 6,9}$	40,7	$52,2 \pm 6,9$	37,9	$53,2 \pm 6,9$	39,4

TABLE A.4 – Matrice de confusion du système à trois rôles obtenue avec les paramètres acoustiques, après AFD et modélisation par une loi Gaussienne.

	présentateur	*journaliste*	*autre*
présentateur	14	12	0
journaliste	7	47	36
autre	3	28	56

TABLE A.5 – Performances du système à 3 rôles utilisant les paramètres temporels seuls : évaluation sur ESTER-tst, (1) sans réduction de dimension, ou après (2) ACP, ou (3) AFD.

	(1) espace initial		(2) après ACP		après AFD	
	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$
Modèle de Gaussienne	$70,4 \pm 6,3$	77,3	$63,0 \pm 6,7$	63,7	$67,5 \pm 6,5$	72,1
k-ppv	$69,5 \pm 6,4$	71,4	$68,0 \pm 6,4$	69,2	$68 \pm 6,4$	73,1
SVM rbf	$67,5 \pm 6,5$	71,7	$66,5 \pm 6,5$	64,8	$66,5 \pm 6,5$	71,3
SVM polynomial	$70,4 \pm 6,3$	74,2	$64,0 \pm 6,6$	57,9	$66,5 \pm 6,5$	74,4
SVM sigmoïdal	$\mathbf{70,9 \pm 6,3}$	78	$\mathbf{73,9 \pm 6,1}$	$\mathbf{80,8}$	$69,5 \pm 6,4$	$\mathbf{76,6}$
SVM linéaire	$69,5 \pm 6,4$	$\mathbf{79,4}$	$67,5 \pm 6,5$	73,8	$\mathbf{70,0 \pm 6,3}$	76,0

TABLE A.6 – Matrice de confusion du système à trois rôles obtenue avec les paramètres temporels, après ACP et classification par SVM à noyau sigmoïdal, testé sur ESTER-tst.

	présentateur	*journaliste*	*autre*
présentateur	25	1	0
journaliste	0	68	22
autre	2	28	57

A.2.3 Paramètres prosodiques

La table A.7 rassemble les performances atteintes avec l'utilisation des paramètres prosodiques seuls. La table A.8 présente la matrice de confusion pour le meilleur système à trois rôles fondée sur l'utilisation des paramètres prosodiques seuls.

TABLE A.7 – Performances du système à 3 rôles utilisant les paramètres prosodiques seuls : évaluation sur ESTER-tst, (1) sans réduction de dimension, ou après (2) ACP, ou (3) AFD.

	(1) espace initial		(2) après ACP		(3) après AFD	
	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$	$\sigma(\%)$	$\tau(\%)$
Modèle de Gaussienne	$63,1 \pm 6,6$	67,1	$68 \pm 6,4$	68,3	$70 \pm 6,3$	71,9
k-ppv	$65,5 \pm 6,5$	58,6	$72,4 \pm 6,2$	68,9	$\mathbf{75,9 \pm 5,9}$	72
SVM rbf	$75,9 \pm 5,9$	**82,9**	$\mathbf{81,8 \pm 5,3}$	**92,1**	$75,4 \pm 5,9$	**80,1**
SVM polynomial	$\mathbf{76,4 \pm 5,8}$	82,7	$78,8 \pm 5,6$	89,3	$73,4 \pm 6,1$	**80,1**
SVM sigmoïdal	$73,4 \pm 6,1$	81,4	$72,9 \pm 6,1$	79,6	$69,5 \pm 6,3$	77,8
SVM linéaire	$71,9 \pm 6,2$	77,5	$70,4 \pm 6,3$	73,5	$68 \pm 6,4$	73

TABLE A.8 – Matrice de confusion du système à trois rôles obtenue avec les paramètres prosodiques, après ACP et classification par SVM à noyau gaussien rbf.

	présentateur	*journaliste*	*autre*
présentateur	25	1	0
journaliste	1	78	11
autre	7	17	63

Notons que les paramètres prosodiques ont également un fort impact, similaire à celui des paramètres temporels sur la différenciation des locuteurs *présentateur* par rapport aux deux autres catégories. Nous constatons également que leur impact sur la distinction entre la catégorie *journaliste* et *autre* est bien plus fort que celui des paramètres temporels seuls.

A.2.4 Paramètres temporels et prosodiques

La table A.9 rassemble les performances atteintes avec l'utilisation simultanée des paramètres temporels et prosodiques. La table A.10 présente la matrice de confusion pour le meilleur système à trois rôles fondée sur l'utilisation des paramètres temporels et prosodiques.

TABLE A.9 – Performances du système à 3 rôles utilisant les paramètres temporels et prosodiques : évaluation sur ESTER-tst, (1) sans réduction de dimension, ou après (2) ACP, ou (3) AFD.

	(1) espace initial		(2) après ACP		(3) après AFD	
	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$
Modèle de Gaussienne	$71,9 \pm 6,2$	76,8	$69,5 \pm 6,4$	64,4	$72,9 \pm 6,1$	76,4
k-ppv	$68,0 \pm 6,4$	58,3	$68,5 \pm 6,4$	67,2	$\mathbf{76,4 \pm 5,9}$	79,2
SVM rbf	$81,3 \pm 5,4$	84,9	$80,8 \pm 5,4$	82,8	$74,4 \pm 6,0$	79,4
SVM polynomial	$\mathbf{82,3 \pm 5,3}$	84,7	$\mathbf{82,3 \pm 5,3}$	**84,0**	$73,4 \pm 6,1$	79,1
SVM sigmoïdal	$80,8 \pm 5,4$	**87,8**	$75,4 \pm 5,9$	78,4	$75,4 \pm 5,9$	**81,0**
SVM linéaire	$79,8 \pm 5,5$	84,9	$76,4 \pm 5,6$	79,3	$74,9 \pm 6,0$	80,4

TABLE A.10 – Matrice de confusion du système à trois rôles obtenue avec les paramètres temporels et prosodiques classés dans l'espace initial par des SVM à noyau polynomial.

	présentateur	*journaliste*	*autre*
présentateur	23	2	1
journaliste	0	82	8
autre	0	25	62

Notons que la prise en compte des paramètres temporels et prosodiques simultanément a un impact un peu plus faible que chaque type de paramètres pris isolément. Cette association renforce également la distinction entre la catégorie *journaliste* et *autre* et surtout diminue la confusion entre *autre* et *présentateur*.

A.2.5 Synthèse des performances du systèmes à trois rôles

La table A.11 présente une synthèse des performances atteintes avec tous les sous-ensembles de paramètres considérés tour à tour.

A.3 Distinction entre locuteurs ponctuels - non ponctuels

La table A.12 rassemble les performances atteintes avec la définition des 5 rôles prenant en compte les ponctuels et les non ponctuels La matrice de confusion pour le système don-

nant la meilleure performance avec la distinction ponctuels/non ponctuels est présentée dans la table A.13.

TABLE A.11 – Caractéristiques des variantes du système à trois rôles ayant donné les meilleures performances pour chaque jeu de paramètres.

paramètres	métode de classif.	réd. de dim.	nb de dim.	$TRC \pm I_{95\%}$	τ
acoustiques	Mod. de Gaussienne	AFD	2	$57,6\% \pm 6,8$	56,1%
temporels	SVM sigmoïdal	ACP	8	$73,9\% \pm 6,1$	80,8%
prosodiques	SVM rbf	ACP	9	$81.8\% \pm 5,3$	$92,1\%$
pros. + temp.	SVM polynomial	-	26	$82,3\% \pm 5,3$	$84,7\%$
tous	SVM rbf	AFD	2	$79,8\% \pm 5,5$	$85,7\%$

TABLE A.12 – Performances du système à 5 rôles : évaluation sur ESTER-tst, (1) sans réduction de dimension, ou après (2) ACP, ou (3) AFD.

	(1) espace intial		(2) après ACP		(3) après AFD	
	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$	$TRC(\%)$	$\tau(\%)$
Modèle de Gaussienne	$75,4 \pm 5,9$	83,7	$67,5 \pm 6,5$	63,7	$72,4 \pm 6,2$	77,9
k-ppv	$61,1 \pm 6,5$	62,6	$68,0 \pm 6,5$	61,5	$71,4 \pm 6,2$	78,2
SVM rbf	$77,3 \pm 5,8$	79,7	$73,9 \pm 6,1$	65,0	$67,0 \pm 6,5$	69,6
SVM polynomial	$77,3 \pm 5,8$	77,2	$69,5 \pm 6,3$	66,6	$70,4 \pm 6,3$	78,1
SVM sigmoïdal	$\mathbf{81,8 \pm 5,4}$	**88,6**	$70,9 \pm 6,2$	78,1	$\mathbf{70,4 \pm 6,3}$	**79,1**
SVM linéaire	$78,8 \pm 5,6$	82,4	$\mathbf{77,8 \pm 5,7}$	**83,7**	$70,4 \pm 6,3$	76,5

TABLE A.13 – Matrice de confusion du système à 5 rôles obtenue avec distinction ponctuels/non-ponctuels, dans l'espace initial avec SVM à noyau polynomial, testé sur ESTER-tst.

	présentateur	*journaliste non ponctuel*	*autre non ponctuel*	*journaliste ponctuel*	*autre ponctuel*
présentateur	24	2	0	-	-
journaliste non ponctuel	2	47	6	-	-
autre non ponctuel	0	19	40	-	-
journaliste ponctuel	-	-	-	30	5
autre ponctuel	-	-	-	3	25

Nous constatons que la confusion entre *journaliste* et *autre* est plus réduite dans la branche « locuteurs ponctuels » que « non ponctuels ». Cette distinction n'a pas d'impact négatif sur la classification des locuteurs *présentateur*.

A.4 Système hiérarchique avec sélection de paramètres de type RSE sur ESTER2

Cette section donne le détail de chaque étape de classification à deux classes proposée dans le système hiérarchique à cinq rôles. Les variantes décrites concernent 4 méthodes de classification. Les résultats sont obtenus après une étape de sélection de paramètres (RSE). Les tests sont effectués également sur ESTER-tst à partir de segmentations automatiques en locuteurs.

A.4.1 Classification présentateur / non présentateur

La classification à l'aide des SVM linéaires réalise 92% de classification correcte en ayant conservé 14 paramètres : 5 paramètres temporels, 3 paramètres acoustiques et 6 paramètres prosodiques.

TABLE A.14 – Performances du système hiérarchique à 5 rôles pour la classification présentateur/non présentateur précédée d'une étape de sélection de paramètres RSE et appliquée à ESTER-tst : (1) nombre de paramètres conservés et (2) TRC.

présentateur / non présentateur	(1) nb de paramètres	(2) *TRC*
Modèle de Gaussienne	**36**	$\mathbf{96,4 \pm 3,1}$
k-ppv (k=3)	8	$86,4 \pm 5,7$
SVM rbf	17	$81,4 \pm 6,5$
SVM linéaire	14	$92,1 \pm 4,5$

A.4.2 Classification journaliste non ponctuel / autre non ponctuel

Le système hiérarchique à cinq rôles, et notamment dans sa branche dédiée aux locuteurs non pontuels, donne les résultats présentés dans la table A.15.

TABLE A.15 – Performances du système hiérarchique à 5 rôles pour la classification journaliste non ponctuel/autre non ponctuel, précédée d'une étape de sélection de paramètres RSE et appliquée à ESTER-tst : (1) nombre de paramètres conservés et (2) TRC.

journaliste non ponctuel / autre non ponctuel	(1) nb de paramètres	(2) *TRC*
Modèle de Gaussienne	34	$68,4 \pm 8,6$
k-ppv (k=17)	22	$68,4 \pm 8,6$
SVM rbf	30	$63,2 \pm 8,9$
SVM linéaire	**26**	$\mathbf{74,6 \pm 8}$

L'ensemble des paramètres conservés se compose de la totalité des 14 paramètres temporels, de 4 paramètres acoustiques et des 8 paramètres prosodiques. Les paramètres prosodiques conservés sont :

- $\overline{F_0}$ la valeur moyenne du pitch,
- T_{vois} le taux de zones voisées,
- N_{voy} le nombre total de voyelles,
- $Debit_{voy}$ le débit de voyelles ,
- N_{sil} le nombre total de silences,
- $Debit_{sil}$ le débit de silences,
- $\overline{Duree_{sil}}$ la durée moyenne des silences,
- $var(Duree_{sil})$ la variance de la durée des silences.

A.4.3 Classification autre ponctuel / journaliste ponctuel

Le système hiérarchique à cinq rôles, et notamment dans sa branche dédiée aux locuteurs pontuels, donne les résultats présentés dans la table A.16.

TABLE A.16 – Performances du système hiérarchique à 5 rôles pour la classification journaliste ponctuel/autre ponctuel, précédée d'une étape de sélection de paramètres RSE et appliquée à ESTER-tst : (1) nombre de paramètres conservés et (2) TRC.

journaliste ponctuel / autre ponctuel	(1) nb de paramètres	(2) TRC
Modèle de Gaussienne	20	$80,9 \pm 9,8$
k-ppv (k=13)	9	$49,2 \pm 12,4$
SVM rbf	21	$85,7 \pm 8,7$
SVM linéaire	**10**	$\mathbf{88,9 \pm 7,8}$

À noter : plusieurs remarques peuvent être faites. La sélection de paramètres couplée à l'algorithme des k-ppv a généré vraisemblablement un sur-apprentissage des paramètres. Seuls 9 paramètres (3 paramètres acoustiques et 6 paramètres prosodiques) ont été conservés et le taux de reconnaissance est très faible avec seulement 49% de rôles bien reconnus.

Le meilleur système correspond est un SVM linéaire avec lequel le taux de reconnaissance atteint 88,9%. Dans cette configuration 10 paramètres sont conservés. Il s'agit :

- du paramètre temporel : A l'activité globale,
- de 3 paramètres acoustiques : P, $min(P_{loc})$ et $max(P_{loc})$ qui sont respectivement la puissance du signal, les valeurs minimales et maximales de la puissance du signal sur les zones attribuées au locuteur,
- ainsi que 6 paramètres prosodiques : $\overline{F_0}$, T_{vois}, N_{voy}, N_{sil}, $\overline{Duree_{sil}}$ et $var(Duree_{sil})$ qui sont respectivement la valeur moyenne du pitch, le taux de zones voisées, le nombre total de voyelles et de silences, la valeur moyenne et la variance sur la durée des silences.

A.5 EPAC

A.5.1 L'approche générative contre l'approche discriminative

La classification est conduite dans un premier temps sans appliquer de méthode de réduction de dimension. Nous utilisons la même méthode de classification à toutes les étapes de l'architecture hiérarchique.

Le GMM, appris sur les classes *journaliste ponctuel* et *autre ponctuel*, compte 2 composantes. À l'étape de classification *présentateur* contre « non présentateur », le GMM se compose de 8 gaussiennes. Les modèles pour les classes *journaliste non ponctuel* et *autre non ponctuel* ont été appris avec 4 gaussiennes chacun.

Les taux de reconnaissance correcte sont les suivants :

- $TRC = 74,8\% \pm 8,26$, pour l'approche GMM : la matrice de confusion correspondante est indiquée sur la table A.17,
- $TRC = 88,9\% \pm 6,00$, pour le classifieur SVM linéaire : la matrice de confusion est indiquée dans la table A.18.

TABLE A.17 – Matrice de confusion sur le corpus EPAC, en utilisant uniquement des GMM.

	présentateur	*journaliste non ponctuel*	*autre non ponctuel*	*journaliste ponctuel*	*autre ponctuel*
présentateur	12	1	0	-	-
journaliste non ponctuel	2	7	1	-	-
autre non ponctuel	4	14	21	-	-
journaliste ponctuel	-	-	-	35	1
autre ponctuel	-	-	-	4	5

TABLE A.18 – Matrice de confusion sur le corpus EPAC, en utilisant uniquement des SVM linéaires.

	présentateur	*journaliste non ponctuel*	*autre non ponctuel*	*journaliste ponctuel*	*autre ponctuel*
présentateur	12	0	1	-	-
journaliste non ponctuel	1	8	1	-	-
autre non ponctuel	5	3	31	-	-
journaliste ponctuel	-	-	-	35	1
autre ponctuel	-	-	-	0	9

Plus particulièrement, nous observons sur les matrices de confusion que la classe *présentateur* est reconnue de manière similaire par les deux approches. La matrice pour l'approche GMM

montre une confusion importante entre les classes *journaliste non ponctuel* et *autre non ponctuel*. La confusion entre ces deux classes est considérablement réduite par l'utilisation d'un classifieur SVM linéaire. La classification *journaliste ponctuel* contre *autre ponctuel* est également meilleure avec la méthode de classification discriminative.

La différence de performances entre ces deux méthodes de classification est statistiquement significative. Nous poursuivrons nos investigations sur le corpus EPAC, en utilisant le classifieur SVM linéaire et en le combinant à des méthodes de réduction de la dimension

A.5.2 Classifieur SVM linéaire et AFD

L'architecture hiérarchique ramène chaque étape de classification à un problème à deux classes. L'analyse factorielle discriminante projette les données dans un espace de dimension égale au nombre de classes moins une. La classification est ainsi réalisé dans un espace monodimensionnel.

Le taux de reconnaissance global atteint $88,8\% \pm 6,00$. Ce résultat est similaire au TRC obtenu sans méthode de réduction de la dimension, avec la même méthode de classification.

Nous avons identifié les paramètres les plus discriminant formant l'espace des paramètres transformé à chaque étape de réduction de dimensionnalité. Il s'agit :

- pour les classes *journaliste ponctuel* contre *autre ponctuel* des paramètres suivant : A, $\overline{P}$ et $min(P_{loc})$ qui sont respectivement l'activité totale, la puissance moyenne du signal et la puissance minimale du signal sur les zones attribuées au locuteur,
- dans l'étape de classification *présentateur* contre « non présentateur », les paramètres sont : $\overline{I_{seg}}$, $min(I_{seg})$, $\overline{P}$, $min(P)$, $min(P_{loc})$ qui sont respectivement l'inter-segment moyen, l'inter-segment minimum, la puissance moyenne du signal, la puissance minimale du signal et la puissance minimale du signal sur les zones attribuées au locuteur.
- l'AFD appliquée avant l'étape de classification *journaliste non ponctuel* contre *autre non ponctuel* met en relief les paramètres suivant : $\overline{P}$, $min(P_{loc})$ qui sont la puissance moyenne du signal et la puissance minimale du signal sur les zones de parole attribuées au locuteur.

Ces résultats révèlent que sur le corpus EPAC contrairement au corpus ESTER2, les paramètres acoustiques comme $\overline{P}$ *la puissance moyenne du signal* et $min(P_{loc})$ *la valeur minimale de la puissance du signal sur les zones attribuées au locuteur* sont des paramètres avec un pouvoir de séparation important pour les couples de rôles considérés. La matrice de confusion pour ce système est présentée dans la table A.19. Comparativement au système précédent, l'AFD réduit légèrement la qualité de la reconnaissance et introduit notamment des erreurs supplémentaires lors la reconnaissance du rôle *présentateur*.

A.5.3 Classifieur SVM linéaire et sélection de paramètres par Recherche Séquentielle par Élimination

Nous réalisons à présent une sélection de paramètres avant chaque étape de classification.

TABLE A.19 – Matrice de confusion sur le corpus EPAC, en utilisant le SVM linéaire combiné à l'AFD.

	présentateur	*journaliste non ponctuel*	*autre non ponctuel*	*journaliste ponctuel*	*autre ponctuel*
présentateur	10	3	0	-	-
journaliste non ponctuel	1	8	1	-	-
autre non ponctuel	4	1	34	-	-
journaliste ponctuel	-	-	-	35	1
autre ponctuel	-	-	-	1	8

La sélection de paramètres appliquée aux rôles *journaliste ponctuel* et *autre ponctuel* a réduit l'ensemble des paramètres à :

- 1 paramètre temporel A l'activité globale,
- 4 paramètres acoustiques : $var(P)$, $\overline{P_{env}}$, $\overline{P_{loc}}$ et $min(P_{loc})$,
- 6 paramètres prosodiques : $var(F_0)$, $Debit_{voy}$, N_{sil}, $Debit_{sil}$, $\overline{Duree_{sil}}$, $var(Duree_{sil})$.

Pour l'étape *présentateur* contre « non présentateur » :

- 3 paramètres temporels : NsE, $var(I_{seg})$, $min(I_{seg})$,
- 8 paramètres acoustiques : $\overline{P}$, $var(P)$, $\overline{P_{env}}$, $var(P_{env})$ $max(P_{env})$, $min(P_{env})$, $\overline{P_{loc}}$, $min(P_{loc})$,
- 6 paramètres acoustiques : $\overline{F_0}$, $Debit_{voy}$, N_{sil}, $Debit_{sil}$, $\overline{Duree_{sil}}$, $var(Duree_{sil})$.

Pour l'étape *journaliste non ponctuel* contre *autre non ponctuel* :

- 5 paramètres temporels : A, NsA, $\overline{L_{seg}}$, $var(L_{seg})$, $min(L_{seg})$,
- 5 paramètres acoustiques : $\overline{P_{env}}$, $min(P_{env})$, $var(P_{loc})$, $max(P_{loc})$, $min(P_{loc})$,
- 4 paramètres prosodiques : T_{vois}, $\overline{Duree_{voy}}$, $Debit_{sil}$, $var(Duree_{sil})$.

Nous observons à nouveau l'importance des paramètres acoustiques dans l'étape de réduction de dimension. En particulier, lors de la classification de la classe *présentateur* où 8 paramètres sont conservés.

Le taux de reconnaissance correcte est de 83,2% $\pm 7,12$. Ce score représente un chute de près de 6% par rapport au système sans réduction de la dimension. Les erreurs de confusion entre les différentes classes sont présentées dans la table A.20. La confusion est globalement plus importante pour toutes les classes de rôles. Par exemple, le rôle *présentateur* est beaucoup moins bien reconnu par rapport aux configurations précédentes. À nouveau, nous constatons que l'approche utilisant une sélection de paramètres séquentielle par élimination ne converge pas vers un ensemble de paramètres pertinents pour la phase de reconnaissance.

TABLE A.20 – Matrice de confusion sur le corpus EPAC en utilisant le système hiérarchique à cinq rôles avec sélection de paramètres (RSE).

	présentateur	*journaliste non ponctuel*	*autre non ponctuel*	*journaliste ponctuel*	*autre ponctuel*
présentateur	8	1	4	-	-
journaliste non ponctuel	0	4	6	-	-
autre non ponctuel	2	0	37	-	-
journaliste ponctuel	-	-	-	31	5
autre ponctuel	-	-	-	0	9

A.5.4 Classifieur SVM linéaire et ACP

Nous intégrons au système une réduction de la dimension par analyse en composantes principales. La réduction de dimension est achevée en conservant les composantes principales correspondant à 99% de la variance initiale des données :

- la dimension passe de 36 paramètres à 19 composantes pour la classification *présentateur* contre « non présentateur »,
- de la même manière 19 composantes sont conservées lors de la réduction de dimension appliquée aux classes *journaliste non ponctuel* et *autre non ponctuel*,
- pour les classes *journaliste ponctuel* et *autre ponctuel* la dimension passe de 25 paramètres à 13 composantes.

Le taux de reconnaissance correcte obtenu pour ce système est égal à $92\% \pm 5,28$.

La matrice de confusion correspondante est présenté dans la table A.21.

TABLE A.21 – Matrice de confusion sur le corpus EPAC en utilisant le système hiérarchique à cinq rôles avec réduction de dimension (ACP).

	présentateur	*journaliste non ponctuel*	*autre non ponctuel*	*journaliste ponctuel*	*autre ponctuel*
présentateur	12	0	1	-	-
journaliste non ponctuel	1	7	2	-	-
autre non ponctuel	4	0	35	-	-
journaliste ponctuel	-	-	-	35	1
autre ponctuel	-	-	-	0	9

Nous pouvons y observer que la confusion entre les classes *journaliste non ponctuel* et *autre ponctuel* a été considérablement réduite. La classification entre *présentateur* et *autre non ponctuel* s'est également légèrement réduite.

L'ajout d'une analyse en composantes principales en amont de l'étape de classification a permis un gain de 3% sur les performances de classification tout en réduisant la dimension de l'espace des paramètres. Cette amélioration n'est toutefois pas statistiquement significative, au regard de l'intervalle de confiance à 95%.

Annexe B

Format des fichiers structurés fournis dans le cadre du projet EPAC

Formats d'échange des données du SP6 d'EPAC

1 Contexte

Ce document définit le format d'échange des données du Sous-Projet 6 (SP6) du projet EPAC.

1.1 Objectifs

Les objectifs principaux du SP6 sont :
- la mise en évidence des zones de parole conversationnelle,
- l'extraction automatique de la structure d'un document,
- la constitution de collections par regroupement de documents ayant des structurelles temporelles similaires.

1.2 Données d'entrée

Les données d'entrée du SP6 sont actuellement les résultats du SP2 : étiquetages issus de l'extraction de caractéristiques acoustiques dites « bas niveau » (segmentations et regroupements en locuteur, segmentations en événements sonores, etc.). Par la suite, tout résultat de type segmentation temporelle pourra être utilisé.

1.3 Extraction automatique des zones d'interaction

Nous recherchons les zones qui, du point de vue de leurs structures temporelles, semblent contenir des échanges verbaux entre plusieurs locuteurs. En effet, ces zones sont susceptibles de contenir de la parole conversationnelle c'est-à-dire ayant un indice de spontanéité assez élevé.

L'extraction des zones d'interaction est une première étape vers l'extraction automatique des zones de parole conversationnelle.

Elle est réalisée à partir des segmentations en locuteurs. Celles-ci nous sont fournies au format « mdtm » par le SP2 (dont est extraite la table 1). Elles contiennent les informations temporelles sur l'activité en parole des locuteurs d'un même document : la première colonne indique le nom du fichier source, les $3^{ème}$ et $4^{ème}$ colonnes indiquent respectivement le début du segment et sa durée tandis que la dernière colonne porte l'étiquette asscociée au locuteur.

TABLE 1 – Extrait d'un fichier de SRL au format « mdtm ».

...							
20040920_1255_1335_INTER_ELDA	1	502.170	11.400	speaker	NA	U	S3
20040920_1255_1335_INTER_ELDA	1	513.570	70.390	speaker	NA	U	S7
20040920_1255_1335_INTER_ELDA	1	583.960	6.130	speaker	NA	U	S2
20040920_1255_1335_INTER_ELDA	1	590.090	0.780	speaker	NA	U	S4
20040920_1255_1335_INTER_ELDA	1	590.870	9.470	speaker	NA	U	S2
20040920_1255_1335_INTER_ELDA	1	600.340	37.420	speaker	NA	U	S4
20040920_1255_1335_INTER_ELDA	1	637.760	6.060	speaker	NA	U	S2
20040920_1255_1335_INTER_ELDA	1	643.820	35.670	speaker	NA	U	S4
20040920_1255_1335_INTER_ELDA	1	679.490	16.960	speaker	NA	U	S2
20040920_1255_1335_INTER_ELDA	1	696.450	70.920	speaker	NA	U	S4
20040920_1255_1335_INTER_ELDA	1	767.370	0.020	speaker	NA	U	S7
20040920_1255_1335_INTER_ELDA	1	767.390	8.850	speaker	NA	U	S7
20040920_1255_1335_INTER_ELDA	1	776.240	57.930	speaker	NA	U	S4
...							

Sur la figure 1, nous pouvons suivre, une représentation des interventions de deux locuteurs. Cette représentation binarisée (locuteur actif = 1 ; non actif = 0) utilise la seconde comme unité temporelle.

Nous constatons que les représentations des segmentations de ces deux locuteurs sont très différentes. Le *locuteur 1* est beaucoup plus actif que le *locuteur 2* et son activité en parole est également beaucoup plus étendue. L'activité parole du *locuteur* 2 est quant à elle très localisée sur le début du document.

Nous pouvons voir que plusieurs segments de parole du *locuteur 1* sont temporellement très proches des segments du *locuteur 2*. Dans l'étape suivante, nous tentons de mettre en évidence et de caractériser l'interactivité de deux locuteurs dans certaines zones du document.

Les zones d'interaction sont extraites des segmentations en plusieurs locuteurs. Ce sont des zones particulières du document dans lesquelles il existe des alternances de tours de parole entre au moins deux locuteurs.

Le motif générique des zones d'interaction est le suivant :

$$Locuteur_1/Locuteur_2/Locuteur_1/(Locuteur_2/Locuteur_1)^*[Locuteur_2]$$

Ce motif exprime que la plus petite zone d'interaction considérée est de la forme $(Locuteur_1/Locuteur_2/Locuteur_1)$ à laquelle peut venir s'ajouter un certain nombre de fois la séquence $(Locuteur_2/Locuteur_1)$, le motif pouvant également se terminer par un tour de parole du second locuteur $(Locuteur_2)$.

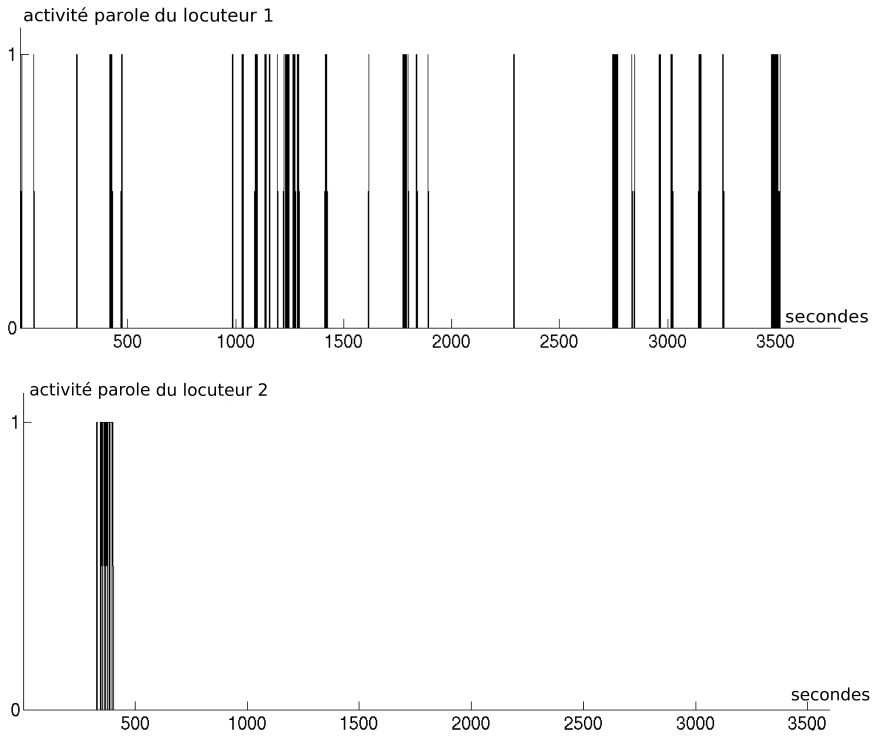


FIGURE 1 – Activité de deux locuteurs sur un document de France Inter d'une heure.

Remarque : l'ordre des indices des locuteurs n'a pas d'incidence et la longueur de la séquence n'est pas limitée.

Dans le corpus sur lequel nous travaillons, un grand nombre de zones d'interaction sont présentes : 20 zones pour une heure d'émission en moyenne. Une grande diversité en terme de longueur et de forme apparaît également. La figure 2 nous montre les zones d'interaction extraites d'une émission d'information de France Inter : les zones d'interaction détectées recouvrent 67% du temps de parole total. Dans d'autres documents, au contraire, il n'y a pas (ou très peu) de zones d'interaction.

Pour distinguer et comparer les zones d'interaction, nous proposons une première mesure simple qui concerne le nombre d'échanges entre les intervenants.

Cette mesure est le niveau d'interactivité. Elle se définit par :

$$niveau\_interactivite = nb\_tour\_parole - 2$$

La zone d'interaction la plus courte considérée est :

$$Locuteur_1/Locuteur_2/Locuteur_1$$

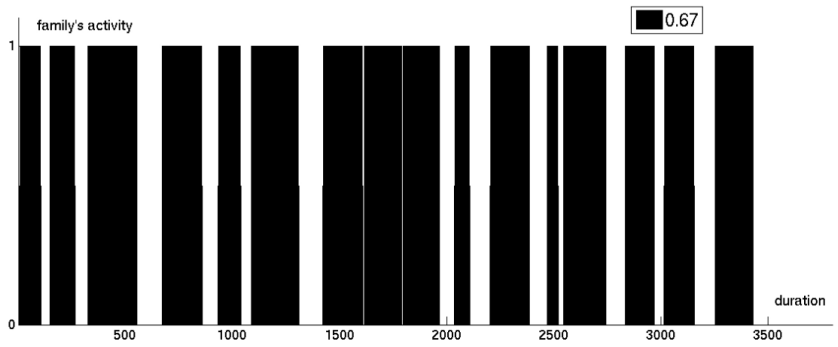


FIGURE 2 – Segmentation en zones d'interaction sur une émission de France Inter.

Son niveau d'interactivité est alors égal à 1.

Remarque : tout segment isolé, c'est-à-dire n'ayant pas été rattaché à une zone d'interaction de niveau au moins égal à 1 (interaction avec un autre locuteur) sera considéré comme une zone d'interaction de niveau 0.

La figure 3 est un exemple de zone d'interaction de niveau égal à 3 détectée entre deux locuteurs S_1 et S_2.

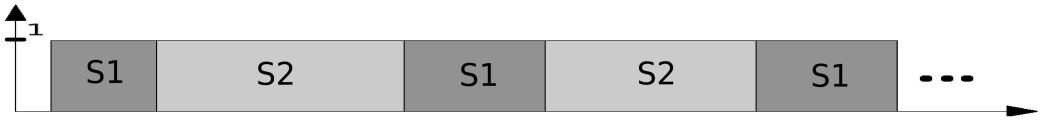


FIGURE 3 – Exemple d'un niveau 3 d'interactivité entre 2 locuteurs.

L'extraction des zones d'interaction est réalisée à partir de segments de parole, sans connaissance *a priori*. Ainsi, les segments sont considérés comme des zones homogènes, contenant de la parole continue (sans silence). Chaque segment est défini par ses indices temporels de début et de fin (fin = début + durée) et l'étiquette correspondant à l'identifiant du locuteur associé au segment. La méthode d'extraction utilisée est basée sur l'analyse de relations temporelles entre segments [Ib2007]. Nous obtnons ainsi une nouvelle segmentation du document en zones d'interaction.

Un problème peut apparaître lorsque 2 zones d'interaction se recouvrent, c'est-à-dire que le dernier segment en locuteur de la première zone peut être également considéré comme le premier segment de la zone d'interaction suivante.

[Ib2007] Zein Al Abidin Ibrahim. *Caractérisation des Structures Audiovisuelles par*

Analyse Statistique des Relations Temporelles. Thèse de doctorat, Université Paul Sabatier, juillet 2007.

Les figures 4 et 5 nous montrent deux zones d'interaction adjacentes. Sur la figure 4, il n'y a pas de recouvrement entre les zones d'interaction (respectivement de niveau 2 et 1). Par contre, sur la figure 5, un segment (Loc 2) est commun aux deux zones d'interaction de niveau 2. Les flèches indiquent le premier segment des zones d'interaction.

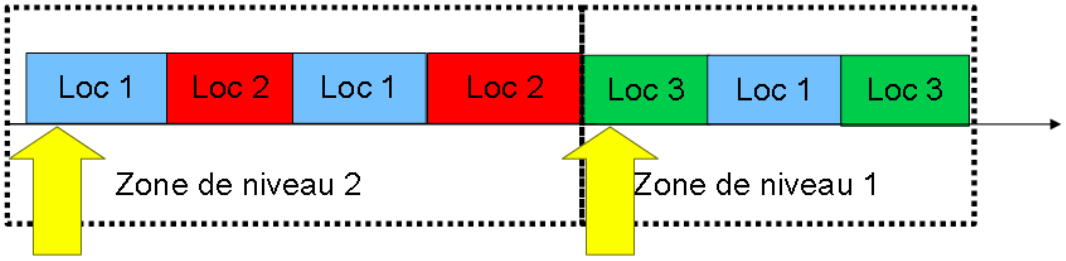


FIGURE 4 – Représentation de 2 zones d'interaction adjacentes et non-recouvrantes.

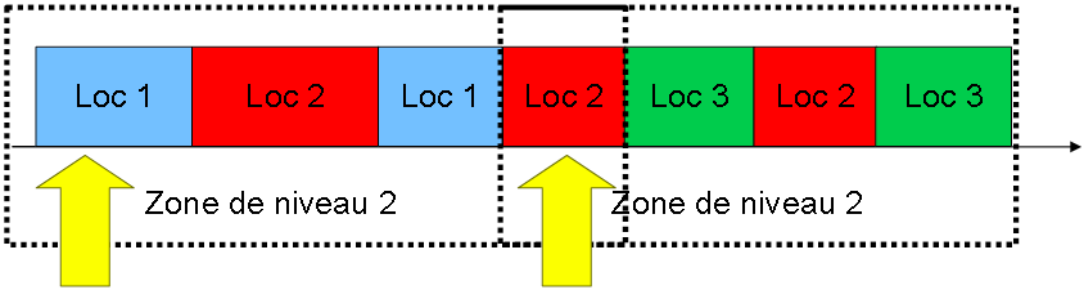


FIGURE 5 – Représentation de 2 zones d'interaction adjacentes et recouvrantes.

Dans le cas d'un recouvrement, trois cas de figure sont alors envisageables :
- soit ce segment appartient aux 2 zones d'interaction,
- soit il appartient à une seule des deux zones
- soit il n'appartient à aucune zone d'interaction.

Quel que soit le cas, il est important de conserver cette information en indiquant qu'il y a un recouvrement potentiel avec une autre zone. En effet à notre stade nous ne sommes pas en mesure de prendre une décision. Ceci pourra être fait par l'exploitation ultérieure du contenu du segment commun, par d'autres SP (SP3 transcription de la parole ou SP4 identification nommées du locuteur) du projet EPAC. Cela permettra d'analyser plus finement le rôle de ce segment et de voir si effectivement il est déconnecté des deux zones, ou bien si il correspond à la clôture d'une interaction avec un premier

locuteur et/ou à l'ouverture d'une interaction avec un second locuteur.

1.4 Typologie des interventions des locuteurs

Le caractère préparé ou bien spontané de la parole peut être lié au type d'intervention de chaque locuteur. En effet, le présentateur d'un journal d'information ne laisse que peu de place à la spontanéité à travers la lecture de fiches ou d'un prompteur, la préparation des questions aux invités, etc. De plus, comme indiqué précédemment, la détermination du types des intervenants donne une information sur la nature de la séquence où il intervient ce qui, par voie de conséquence, permet d'avoir une représentation de la structure globale du document.

Chaque zone d'interaction détectée est donc enrichie en ce sens. Nous associons à chaque locuteur intervenant dans la zone, ses types d'intervention (lorsque celui-ci peut être déterminé). Cette information est pour le moment définie globalement (à l'échelle du document) à partir de :

- **l'activité** : durée totale des zones de parole du locuteur dans le document,
- **l'étendue** : plage qui recouvre la totalité de l'activité du locuteur (entre sa première et sa dernière interventions).

La projection des descripteurs d'activité et d'étendue, permet de représenter chaque locuteur d'un même document (cf. figure 6) par un point. Ceci révèle une répartition particulière des points sur le plan et ce pour un grand nombre d'émissions.

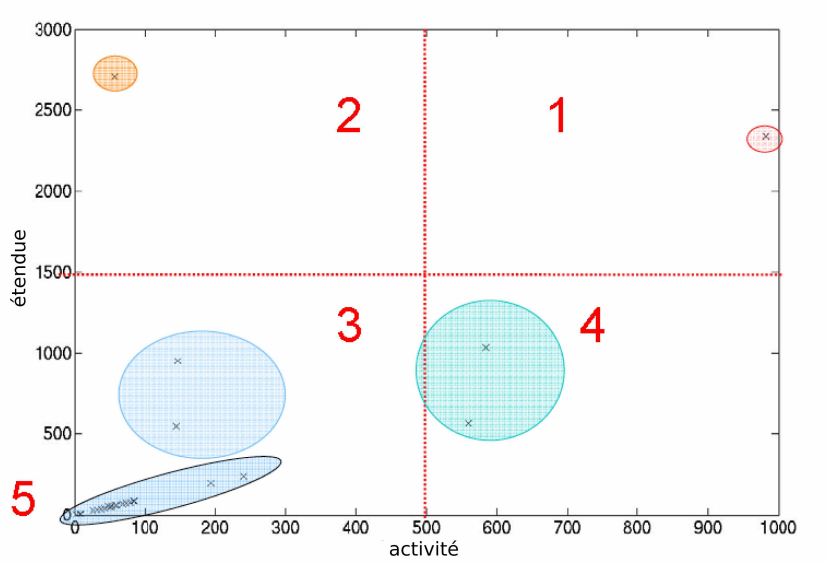


FIGURE 6 – Rôle des locuteurs dans le plan Activité-Etendue.

Ceci nous a conduit à découper le plan Activité-Etendue en 5 sous-sections indicées de 1 à 5, définissant ainsi 5 types d'interventions. Chaque locuteur sera donc caractérisé par l'indice qui lui correspond.

- L'indice 1 est attribué aux locuteurs les plus actifs ayant une étendue importante (quart supérieur droit du plan). C'est typiquement le cas des présentateurs.

- L'indice 2 correspond à des locuteurs dont l'étendue est plus importante que l'activité (quart supérieur gauche). C'est le cas des locuteurs apparaisssant plusieurs fois dans l'émission mais qui ne s'expriment pas très longtemps. Nous retrouvons ce comportement chez les speakers et speakerines qui apparaissent en début et fin d'émission pour marquer la transition avec le programme suivant.

- Les indices 3 et 4 (quart inférieurs gauche et droit du plan) ont une étendue réduite mais se distinguent par leur activité (réduite pour 3 et plus importante pour 4). Dans l'exemple de la figure 6, l'étendue est inférieure à un tiers de la durée du document. L'activité de ces locuteurs est donc assez localisée. Ce profil correspond par exemple, à celui du présentateur d'une chronique.

- L'indice 5 constitue un cas particulier de la catégorie 3. Il concerne les locuteurs qui ont une activité égale à leur étendue. Les points qui représentent ces locuteurs sur le plan sont situés sur la première médiatrice. Ce sont des locuteurs ponctuels, c'est-à-dire qu'ils n'apparaissent qu'une seule fois dans le document. Sous ce rôle, nous trouvons entre autres des reporters, des commentaires de personnalités, des extraits de conférences de presse, des personnes interviewées dans la rue...

L'ensemble des informations que nous pouvons extraire à ce stade est mis en forme suivant le format décrit dans la section suivante.

2 Format de sortie du SP6

2.1 Le fichier de sortie

Dans le but de ne pas surcharger les informations dans le fichier de sortie « trs », chaque SP fournit ses résultats dans des fichiers XML. Le SP6 produit un fichier XML. Il concerne les segmentations en zones d'interaction : il est fourni dans des fichiers portant l'extension **'.zi.xml'**. Ce fichier au format XML, rassemble le résultat de l'extraction automatique de la structure des documents obtenus par agrégation des résultats du SP6 et du SP2.

2.2 Segmentations en zones d'interaction

On y trouve l'entête XML suivante :

<**ZI** system="**IRIT_Interact**" version="**2**" audio_filename="**200409_0455_0535_INTER_ELDA**" souce_type="**speakers_auto**" date="**090511**" type="**AUTO**">

avec :

- **system** : le nom du sytème ayant fourni la segmentation en zones d'interaction,
- **version** : la version du système utilisé,
- **audio_filename** : le nom du document audio source,
- **source_type** : le type de segmentation utilisée en entrée. Par exemple, *speakers* correspond à la tâche SRL mais il peut y avoir également *speech, music, silence, etc.* (tout type de segmentation temporelle). Les segmentations peuvent être obtenues de manière automatique *_auto* ou bien manuelle *_ref* (issues de fichiers de référence),
- **date** : la date de création du fichier **'.zi.xml'**,
- **type** : le type de segmentation en zones d'interaction réalisée ($AUTO$ = automatique , $MANUAL$ = manuelle ou $REVISED$ = pour une version automatique, revue manuellement).

Le corps du fichier contient des informations sur chaque zone d'interaction détectée dans le document et délimitée par une balise <**zi**>. L'exemple suivant décrit deux zones d'interaction consécutives extraites du fichier « .mdtm » donné en exemple précédemment (cf. table 1).

...
<**zi** id="**zi014**" startTime="**513.570**" endTime="**583.960**" level="**0**" speakers="**S7**" roles="**2**"> **others** </**zi**>
<**zi** id="**zi015**" startTime="**583.960**" endTime="**767.370**" level="**6**" speakers="**S2 S4**" roles="**5 1**"> **interview** </**zi**>
...

Chaque balise contient un nombre d'attributs obligatoires auxquels peuvent s'ajouter des attributs optionels :

- **id** : identifiant unique de la zone d'interaction dans le document. Il est attribué chronologiquement au fur et à mesure de la détection des zones d'interaction,

- **startTime** et **endTime** : ce sont les instants de début et de fin de la zone d'interaction (valeurs exprimées en secondes),

- **level** : niveau d'interactivité de la zone. Afin de couvrir la totalité du document, les zones de niveau d'interactivité 0 sont renseignées de la même manière que les autres zones,

- **speakers** : liste des identifiants des locuteurs présents dans la zone d'interaction. Le format d'écriture de ces identifiants *(S+nombre)* est celui utilisé par le SP2 dans la tâche SRL,

- **roles** : liste des indices correspondant au type de l'intervention de chaque locuteur présent dans la zone. Par abus de langage, nous nommons cet attribut « rôle ». Cet attribut est à mettre en parallèle avec la liste **speakers**, de telle sorte que le $n^{ième}$ rôle de la liste **roles** corresponde au n^{ieme} locuteur de la liste **speakers**. Si un rôle n'est pas défini il est signalé par la valeur 0.

Voici d'autres attributs qui conservent, pour l'instant, un caractère optionnel :

- **startSentence et endSentence** : font référence aux transcriptions tokénisées. Ces attributs contiennent les identifiants respectifs de la première phrase (ou *sentence*) et de la dernière phrase prononcées dans la zone d'interaction,

- **covered** : indique si la zone est recouverte par une autre zone d'interaction :
  - "10", s'il y a recouvrement en début de zone,
  - "01" si le recouvrement a lieu à la fin,
  - "11" pour un recouvrement en début et en fin de zone,
  - "00" s'il n'y a pas de recouvrement.

- **confidence** : contient une probabilité, qui représente la mesure de confiance globale de la zone,

Le **contenu de la balise** est une information sur la nature de la zone d'interaction. Pour le moment nous prenons en compte les types suivants : « debate », « interviews », « news », « chronicle », « others », etc.

Bibliographie

[Amaral 03] R. Amaral & I. Trancoso. *Topic Indexing of TV Broadcast News Programs.* In Proc. of 6th international conference on computational processing of the portuguese language, 2003.

[André-Larochebouvy 84] D. André-Larochebouvy. La conversation quotidienne. introduction à l'analyse sémio-linguistique de la conversation. Didier Crédif, 1984.

[Auffret 99] G. Auffret & B. Bachimont. *Audiovisual cultural heritage : From TV and radio archiving to hypermedia publishing.* In Proc. of 3rd European Conference on research and advanced technology for Digital Libraries, 1999.

[Aydin Alatan 01] A. Aydin Alatan, A. N. Akansu & W. Wolf. *Multi-Modal Dialog Scene Detection Using hidden Markov models for Content-Based Multimedia Indexing.* Multimedia Tools and Applications, vol. 14, no. 2, pages 137–151, June 2001.

[Bale 50] R.F. Bale. *A set of categories for the analysis of small group interaction.* American Sociological Review, vol. 15, no. 2, pages 257–263, 1950.

[Banerjee 06] S. Banerjee & A. Rudnicky. *You are what you say : using meeting participant' speech to detect their roles and expertise.* In Proc. of HLT-NAACL Workshop on Analyzing Conversations in Text and Speech, pages 23–30, 2006.

[Baraggioli 08] J.-L. Baraggioli & B. Desnoues. *Indexation : Dewey et Rameau*, mediadix. pôle des métiers du livre. saint-cloud. france. http ://netx.u-paris10.fr/eadmediadix/intro/index.php edition, 2008.

[Barzilay 00] R. Barzilay, M. Collins, J. Hirschberg & S. Wittaker. *The rules behind the roles : identifying Sepaker role in radio Broadcast.* In Proc. of 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, pages 679–684. AAAI Press / The MIT Press, 2000.

[Basu 02] S. Basu. *Conversational Scene Analysis.* PhD thesis, Massachusset Institute of Technology, 2002.

[Bellman 61] R. Bellman. Adaptive control processes : A guided tour. Princeton University Press, 1961.

[Biddle 86] B.J. Biddle. *Recent Developments in Role Theory.* Annual Review of Sociology, vol. 14, pages 67–92, 1986.

[Bigot 08a] B. Bigot & I. Ferrané. *From Audio Content Analysis to Conversational Speech Detection and Characterization.* In Proc. of ACM SIGIR Workshop : Searching Spontaneous Conversational Speech (SSCS), pages 62–65. ILPS, 2008.

[Bigot 08b] B. Bigot, I. Ferrané & Z. A. A. Ibrahim. *Towards the detection and the characterization of conversational speech zones in audiovisual documents.* In Proc. of IEEE International Workshop on Content-Based Multimedia Indexing, 2008.

[Bigot 08c] B. Bigot, I. Ferrané & Z.A.A. Ibrahim. *Caractérisation des zones d'interactivité entre locuteurs : vers la détection de zones de parole conversationnelle pour le projet EPAC.* In Journées d'Etudes sur la Parole (JEP), pages 169–172. AFCP, 2008.

[Bigot 10a] B. Bigot, I. Ferrané & J. Pinquier. *Exploiting speaker segmentations for automatic role detection. An application to broadcast news documents.* In Proc. of IEEE International Workshop on Content-Based Multimedia Indexing, pages 207–212, 2010.

[Bigot 10b] B. Bigot, I. Ferrané, J. Pinquier & R. André-Obrecht. *Detecting individual role using features extracted from speaker diarization results.* Multimedia Tools and Applications, pages 1–23, 2010.

[Bigot 10c] B. Bigot, I. Ferrané, J. Pinquier & R. André-Obrecht. *Speaker role recognition to help spontaneous conversational speech detection.* In Proc. of International Workshop on Searching Spontaneous Conversational Speech, pages 5–10. ACM, 2010.

[Bigot 10d] B. Bigot, J. Pinquier, I. Ferrané & R. André-Obrecht. *Looking for relevant features for speaker role recognition.* In Proc. of Interspeech, pages 1057–1060, 2010.

[Brdiczka 05] O. Brdiczka, J. Maisonnasse & P. Reignier. *Automatic detection of interaction groups.* In Proc. of 7th International Conference on Multimodal Interfaces, pages 32–36. ACM, 2005.

[Burger 02] S. Burger, V. MacLaren & H. Yu. *The ISL Meeting Corpus : The Impact of Meeting Type on Speech Style.* In Proc. of International Conference on Spoken Language Processing, 2002.

[Canseco 05] L. Canseco, L. Lamel & J.-L. Gauvain. *A comparative study using manual and automatic transcriptions for diarization.* In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, pages 415–419, 2005.

[Carreira-Perpiñán 97] M Carreira-Perpiñán. *A review of dimension reduction techniques.* Rapport technique, Dept. of Computer Science University of Sheffield, 1997.

[Chua 04] T.-S. Chua, S.-F. Chang, L. Chaisorn & W. Hsu. *Story boundary detection in large broadcast news video archives : techniques, experience and trends.* In Proc. of 12th ACM international conference on Multimedia, pages 656–659, 2004.

[Corman 94] S. Corman & C. Scott. *A synchronous digital signal processing method for detecting face-to-face organizational communication behavior.* Social Networks, vol. 16, no. 2, pages 163–179, 1994.

[Covell 06] M. Covell, S. Baluja & M. Fink. *Advertisement detection and replacement using acoustic and visual repetition.* In Proc. of IEEE 8th Workshop on Multimedia Signal Processing, pages 461–466, 2006.

[Cunningham 08] P. Cunningham. *Dimension Reduction.* In M. Cord & P. Cunningham, editeurs, Machine Learning Techniques for Multimedia, Cognitive Technologies, pages 91–112. Springer Berlin Heidelberg, 2008.

[de Cheveigné 02] A. de Cheveigné & H. Kawahara. *YIN, a fundamental frequency estimator for speech and music.* Journal of the Acoustic Society of America, 2002.

[Dempster 77] A. Dempster, N. Laird & D. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pages 1–38, 1977.

[Dimitrova 02] N. Dimitrova, S. Jeannin, J. Nesvadba, T. McGee, L. Agnihotri & G. Mekenkamp. *Real time commercial detection using MPEG features.* In Proc. of International Conference on Information Processing and Management of Uncertainty in knowledge-based systems, pages 1–6, 2002.

[Duda 00] R. O. Duda, P. E. Hart & D. G. Stork. Pattern classification. Wiley-Interscience, 2nd edition, 2000.

[Dufour 09] R. Dufour, Y. Estève, P. Deléglise & F. Béchet. *Local and global models for spontaneous speech segment detection and characterization.* In Proc. of IEEE Workshop on Automatic Speech Recognition Understanding, pages 558–561, 2009.

[Eickeler 99] S. Eickeler & S. Muller. *Content-based video indexing of TV broadcast news using hidden Markov models.* Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 6, pages 2997–3000, 1999.

[El Khoury 08] E. El Khoury, C. Sénac & P. Joly. *Unsupervised TV Program Boundaries Detection Based on Audiovisual Features.* In Proc. of International Conference on Visual Information Engineering. IET, 2008.

[El Khoury 09] E. El Khoury, C. Sénac & J. Pinquier. *Improved speaker diarization system for meetings.* In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4097–4100, 2009.

[El Khoury 10] E. El Khoury. *Unsupervised Video Indexing based on Audiovisual Characterization of Persons.* PhD thesis, Université de Toulouse, 2010.

[Estève 07] Y. Estève, S. Meignier, P. Deléglise & J. Mauclair. *Extracting true speaker identities from transcriptions.* In Proc. of Interspeech, 2007.

[Estève 10] Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet & Jérôme Farinas. *The EPAC Corpus : Manual and Automatic Annotations of Conversational Speech in French Broadcast News.* In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias, editeurs, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[Ferman 99] A.M. Ferman & A.M. Tekalp. *Probabilistic Analysis and Extraction of Video Content.* In Proc. of IEEE International Conference on Image Processing, volume 2, pages 91–95, 1999.

[Fisher 36] R. A. Fisher. *The use of multiple measurements in taxonomic problems.* Annals Eugen., vol. 7, pages 179–188, 1936.

[Furui 05] S. Furui, M. Nakamura, T. Ichiba & K. Iwano. *Why is the recognition of spontaneous speech so hard ?* Lectures notes in Computer Science, 2005.

[Galliano 05] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre & G. Gravier. *The ESTER phase II evaluation campaign for the rich transcription of french broadcast news.* In Proc. of Interspeech, 2005.

[Galliano 09] S. Galliano, G. Gravier & L. Chaubard. *The ESTER 2 evaluation campaign for the rich transcription of french radio broadcast.* In Proc. of Interspeech, 2009.

[Gauvain 05a] J-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel & H. Schwenk. *Where Are We In Transcribing French Broadcast News ?* In Proc. of Interspeech, 2005.

[Gauvain 05b] J-L. Gauvain, G. Adda, L. Lamel, F. Lefèvre & H. Schwenk. *Transcription de la parole conversationnelle*. TAL, vol. 45, no. 3, pages 35–47, 2005.

[Guyon 03] I. Guyon & A. Elisseeff. *An introduction to variable and feature selection*. Journal of Machine Learning Research, vol. 3, pages 1157–1182, March 2003.

[Hauptmann 98] A. Hauptmann & M. Witbrock. *Story Segmentation and Detection of Commercials In Broadcast News Video*. In Advances in Digital Librairies Conference, 1998.

[Ibrahim 07] Z. A. A. Ibrahim. *Caractérisation de Structures Audiovisuelles par Analyse Statistique des Relations Temporelles*. PhD thesis, Université Paul Sabatier, Toulouse, 2007.

[Javed 02] O. Javed, S. Khan, Z. Rasheed & M. Shah. *Visual Content Based Segmentation Of Talk Game Shows*. International Journal of Computers and Applications, 2002.

[Kemp 03] T. Kemp, M. Schmidt, M. Westphaland & A. Waibel. *Strategies for automatic segmentation of audio data*. In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[Kerbrat-Orecchioni 95] C Kerbrat-Orecchioni. Les interactions verbales. Editions Armand Colin, 2nd edition, 1995.

[Kijak 03] E. Kijak. *Structuration multimodale des vidéos de sports par modèles stochastiques*. PhD thesis, Université de Rennes 1, 2003.

[Kolluru 07] B. Kolluru & Y. Gotoh. *Speaker role based structural classification of broadcast news stories*. In Proc. of Interspeech, 2007.

[Kral 05] P. Kral, C. Cerisara & J. Kleckova. *Combination of classifiers for automatic recognition of dialog acts*. In Proc. of Interspeech, 2005.

[Lachambre 09] H. Lachambre. *Caractérisation de l'environnement musical dans les documents audiovisuels*. PhD thesis, Université de Toulouse, 2009.

[Le Blouch 09] O. Le Blouch. *Décodage acoustico-phonétique et applications à l'indexation audio automatique*. PhD thesis, Université de Toulouse, 2009.

[Liang 05] L. Liang, H. Lu, Xue X. & Y.-P. Tan. *Program segmentation for TV videos*. In Proc. of IEEE International Symposium on Circuits and Systems, volume 2, pages 1549–1552, 2005.

[Lienhart 97] R. Lienhart, C. Kuhmunch & W. Effelsberg. *On the detection and recognition of television commercials*. In Proc. of IEEE International Conference on Multimedia Computing and Systems, pages 509–516, 1997.

[Liu 06] Y. Liu. *Initial study on automatic identification of speaker role in broadcast news speech*. In Proc. of Human Language Technology Conference of the NAACL, pages 81–84, 2006.

[Louradour 07] J. Louradour. *Noyaux de séquences pour la vérification du locuteur par Machines à Vecteurs de Support*. PhD thesis, Université Paul Sabatier, Toulouse, 2007.

[Lu 98] G. Lu & T. Hankinson. *A technique towards automatic audio classification and retrieval*. In Proc. of International Conference on Signal Processing, volume 2, pages 1142–1145, 1998.

[Luzzati 04] D. Luzzati. *Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané*. In Proc. of Workshop Modélisation pour l'Identification des Langues, pages 13–17, 2004.

[Ma 09] C. Ma, B. Byun, I. Kim & C.-H. Lee. *A detection-based approach to broadcast news video story segmentation.* In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1957–1960, 2009.

[Manson 10] G. Manson. *Délinéarisation automatique des flux de télévision.* PhD thesis, Université de Rennes 1, 2010.

[Martinez 02] J.M. Martinez. *MPEG-7 : overview of MPEG-7 Description Tools, Part 2.* IEEE Multimedia, vol. 9, no. 3, pages 83–93, 2002.

[Maskey 03] S. Maskey & J. Hirschberg. *Automatic Speech Summarization of Broadcast News using Structural Features.* In Proc. of Eurospeech, 2003.

[Maybury 96] M Maybury, M. Merlino & J. Rayson. *Segmentation, content extraction and visualization of broadcast news video using multistream analysis.* In Proc. of ACM International Conference on Multimedia, 1996.

[Meinedo 03] H. Meinedo & J. Neto. *Audio segmentation, classification and clustering in a broadcast news task.* In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[Merlino 97] A. Merlino, D. Morey & M. Maybury. *Broadcast news navigation using story segmentation.* In Proc. of ACM International Conference on Multimedia, pages 381–391, 1997.

[Naturel 07] X. Naturel. *Structuration automatique de flux vidéos de télévision.* PhD thesis, Université de Rennes 1, 2007.

[Pellegrino 00] F. Pellegrino & R. André-Obrecht. *Automatic language identification : an alternative approach to phonetic modelling.* Signal Processing, vol. 80, no. 7, pages 1231–1244, 2000.

[Pinquier 04] J. Pinquier. *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle.* PhD thesis, Université Paul Sabatier, 2004.

[Poli 07] J.-P. Poli. *Structuration automatique de flux télévisés.* PhD thesis, Université Paul Cézanne Aix-Marseille III, 2007.

[Prié 99] Y. Prié. *Modélisation de documents audiovisuels en Strates Interconnectées par les Annotations pour l'exploitation contextuelle.* PhD thesis, Institut National des Sciences Appliquées de Lyon, 1999.

[Pua 04] K. Pua, J. Gauch, S. Gauch & J. Miadowicz. *Real time repeated video sequence identification.* Computer Vision and Image Understanding, vol. 93, no. 3, pages 310–327, 2004.

[Puigt 06] M. Puigt & Y. Deville. *A time-frequency correlation-based blind source separation method for time-delayed mixtures.* In Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 853–856, 2006.

[Quinlan 93] J. R. Quinlan. C4.5 : programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[Rabiner 75] L. Rabiner & M. Sambur. *An Algorithm for Determining the Endpoints of Isolated Utterances.* Rapport technique, Bell System Technology Journal, 1975.

[Ramona 10] M. Ramona. *Classification automatique de flux radiophoniques par Machine à Vecteurs de Support.* PhD thesis, Télécom Paris Tech, 2010.

[Sacks 74] H. Sacks, E. A. Schegloff & G. Jefferson. *A Simplest Systematics for the Organization of Turn-Taking for Conversation.* Language, vol. 50, no. 4, pages 696–735, 1974.

[Sadlier 02] D. Sadlier, S. Marlow, N. O'Connor & N. Murphy. *Automatic TV advertisement detection from MPEG bitstream.* Pattern Recognition, vol. 35, no. 12, pages 2719–2726, 2002.

[Salamin 09] H. Salamin, S. Favre & A. Vinciarelli. *Automatic Role Recognition in Multiparty Recordings : Using Social Affiliation Networks for Feature Extraction.* IEEE Transactions on Multimedia, vol. 11, no. 7, pages 1373–1380, 2009.

[Saraceno 98] C. Saraceno & R. Leonardi. *Identification of story units in audiovisual sequences by joint audio and video processing.* In Proc. of IEEE Intenational Conference on Image Processing, 1998.

[Shriberg 98] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, A. Erringer, M. Gregory, L. Heintzelman, T. Metzler, A. Oduro & T. The. *Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech ?* Language and speech, vol. 41, pages 443–492, 1998.

[Stolcke 99] A. Stolcke, E. Shriberg, D. Hakkani-Tür, G. Tür, Z. Rivlin & K. Sönmez. *Combining Words and Speech Prosody for Automatic Topic Segmentation.* In Proc. of DARPA Broadcast News Transcription and Understanding Workshop, pages 61–64, 1999.

[Stolcke 00] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema & M. Meteer. *Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech.* Computational Linguistics, vol. 26, pages 339–373, 2000.

[Sudaram 02] H. Sudaram. *Segmentation, Structure Detection and Summarization of Multimedia Sequences.* PhD thesis, Columbia University, 2002.

[Troncy 04] R. Troncy. *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels.* PhD thesis, Université Joseph Fourier - Grenoble 1, 2004.

[Vapnik 98] V. Vapnik. Statistical learning theory. John Wiley & Sons, 1998.

[Vinciarelli 07] A. Vinciarelli. *Speakers Role Recognition in Multiparty Audio Recordings Using Social Network Analysis and Duration Distribution Modeling.* IEEE Transactions on Multimedia, vol. 9, no. 6, pages 1215–1226, 2007.

[Wang 08] J. Wang, L. Duan, Q. Liu, H. Lu & J.S. Jin. *A Multimodal Scheme for Program Segmentation and Representation in Broadcast Video Streams.* IEEE Transactions on Multimedia, vol. 10, no. 3, pages 393–408, 2008.

[Wen 99] X. Wen, T. D. Huffmire, H. H. Hu & A. Finkelstein. *Wavelet-based video indexing and querying.* Multimedia Systems, vol. 7, pages 350–358, 1999.

[Weng 07] C.-Y. Weng, W.-T. Chu & J.-L. Wu. *Movie analysis base on role's social network.* In Proc. of IEEE International Conference on Multimedia & Expo, 2007.

[Wolf 97] W. Wolf. *Hidden Markov Model Parsing of Video Programs.* In Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, pages 2609–2611, 1997.

[Wu 02] S. Wu & P. Flach. *Feature selection with labelled and unlabelled data.* In Proc. of Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning, pages 156–167, 2002.

[Wyatt 07] D. Wyatt, T. Choudhury & J. Bilmes. *Conversation detection and speaker segmentation in privacy-sensitive situated speech data*. In Proc. of Interspeech, 2007.

[Zhai 04] Y. Zhai, Z. Rasheed & M. Shah. *Conversation detection in feature films using finite state machines*. In Proc. of International Conference on Pattern Recognition, pages 458–461, 2004.

Doctorat de l'université de Toulouse 3 - Informatique - **Recherche du rôle des intervenants et de leurs interactions pour la structuration de documents audiovisuels** soutenue par **B. BIGOT** le **6 juillet 2011**, Amphi Schwartz - Institut de Mathématiques de Toulouse - Direction : *R. André-Obrecht*

Résumé

Nous présentons un système de structuration automatique d'enregistrements audiovisuels s'appuyant sur des informations non lexicales caractéristiques des rôles des intervenants et de leurs interactions.

Dans une première étape, nous proposons une méthode de détection et de caractérisation de séquences temporelles, nommées « zones d'interaction », susceptibles de correspondre à des conversations.

La seconde étape de notre système réalise une reconnaissance du rôle des intervenants : *présentateur, journaliste* et *autre*. Notre contribution au domaine de la reconnaissance automatique du rôle se distingue en reposant sur l'hypothèse selon laquelle les rôles des intervenants sont accessibles à travers des paramètres « bas-niveau » inscrits d'une part dans l'organisation temporelle des tours de parole des intervenants, dans les environnements acoustiques dans lesquels ils apparaissent, ainsi que dans plusieurs paramètres prosodiques (intonation et débit).

Dans une dernière étape, nous combinons l'information du rôle des intervenants à la connaissance des séquences d'interaction afin de produire deux niveaux de description du contenu des documents. Le premier niveau de description segmente les enregistrements en zones de 4 types : informations, entretiens, transition et intermède. Un second niveau de description classe les zones d'interaction orale en 4 catégories : débat, interview, chronique et relais. Chaque étape du système est validée par une grand nombre d'expériences menées sur le corpus du projet EPAC et celui de la campagne d'évaluation ESTER.

Mots-clés: structuration de documents audiovisuels ; reconnaissance automatique du rôle ; détection de zones de conversations ; reconnaissance des formes ; paramètres temporels, acoustiques et prosodiques

Abstract

We present a system for audiovisual document structuring, based-on speaker role recognition and speech interaction zone detection.

The first stage of our system consists in an automatic method for speech interaction zones detection and characterization. Such zones correspond to temporal sequences of documents which potentially contain conversations between speakers.

The second stage of our system achieves the recognition of speaker roles : *anchorman, journalist* and *other*. Our contribution to this domain is based on the hypothesis that cues about speaker roles are available through low-level features extracted from the temporal organization of turn-takings and from acoustic and prosodic features (speech rate and pitch).

In the last stage of our system, we combine speaker roles and speech interaction zones to provide two descriptive layers of the audiovisual document contents. The first descriptive layer gathers segments of 4 types : informations, meeting, transition and interlude. The second descriptive layer consists in a classification of speech interaction zones into 4 categories : debate, interview, chronicle and relay. Each step of the system has been evaluated using a large number of experiments realized using the EPAC project and ESTER campaign corpora.

Keywords: audiovisual document structuring ; speaker role recognition ; conversation detection ; pattern recognition ; temporal, acoustic and prosodic features