# Segmentations sonore et audiovisuelle ?

Julien Pinquier

# Mémoire

**En vue de l'obtention de l'**

## HABILITATION A DIRIGER LES RECHERCHES

Délivrée par l'***Université Toulouse III - Paul Sabatier***

Spécialité : ***Informatique***

---

Présentée par **Julien Pinquier**

Le 19 mai 2014

# Segmentations sonore et audiovisuelle ?

---

## JURY

*Rapporteurs*

Laurent BESACIER – Professeur, Université Grenoble 1
Patrick GROS – Directeur de Recherche, INRIA Rennes
Bernard MERIALDO – Professeur, EURECOM

*Examinateurs*

Régine ANDRE-OBRECHT – Professeur, Université Toulouse III
Gaël RICHARD – Professeur, TELECOM ParisTech

*À Nookie,*

# Table des matières

**Introduction générale** **1**

1 Contexte général . . . . . 1
2 Contexte scientifique et organisation du document . . . . . 2

**Partie I Segmentation sonore**

**Introduction** **7**

**Chapitre 1**
**Segmentation de la composante parole** **11**

1.1 Segmentation en locuteurs . . . . . 11
1.1.1 Système SRL initial . . . . . 12
1.1.2 Système SRL enrichi . . . . . 12
1.2 Détection de zones d'interaction . . . . . 13
1.2.1 Contexte et définitions d'une unité d'interaction . . . . . 13
1.2.2 Système de détection et de caractérisation des Z.I. . . . . 14
1.3 Reconnaissance du rôle . . . . . 17
1.3.1 Définition des rôles . . . . . 17
1.3.2 Paramètres « bas-niveau » . . . . . 18
1.3.3 Système de reconnaissance des rôles . . . . . 19

**Chapitre 2**
**Segmentation de la composante musique** **23**

2.1 Discrimination monophonique/polyphonique . . . . . 24
2.1.1 Système « monopoly » . . . . . 24
2.1.2 Performance . . . . . 26
2.2 Détection du chant . . . . . 26
2.2.1 Vibrato . . . . . 27
2.2.2 Système de détection du chant . . . . . 29
2.3 Détection du chœur à l'unisson . . . . . 30
2.3.1 Introduction . . . . . 30
2.3.2 Système de détection du chœur à l'unisson . . . . . 30

**Chapitre 3**
**Au delà des segmentations primaires** **35**

3.1 Segmentation mixte : parole et musique . . . . . 35
3.1.1 Détection de zones superposées . . . . . 35
3.1.1.1 Introduction . . . . . 36
3.1.1.2 Système de détection de sources multiples . . . . . 37
3.1.2 Fusion PMB/SRL . . . . . 39
3.2 Segmentation de la composante bruit . . . . . 41
3.2.1 Détection d'applaudissements et de rires . . . . . 41
3.2.2 Détection de sons d'eau . . . . . 43
3.2.2.1 Système de détection du flot d'eau . . . . . 44
3.2.2.2 Système de détection de gouttes d'eau . . . . . 46

**Discussion** **49**

## Partie II Segmentation audiovisuelle


**Introduction** **55**

**Chapitre 4**
**Segmentation autour des intervenants** **57**

4.1 Détection des intervenants . . . 58
4.1.1 Principe . . . 58
4.1.2 Matrice de co-occurrences . . . 59
4.2 Caractérisation des intervenants . . . 61
4.2.1 Hypothèses de travail . . . 62
4.2.2 Changement d'environnement sonore . . . 62
4.2.3 Système de classification IN-OUT-OFF . . . 63
4.3 Structuration en programmes . . . 65
4.3.1 Structurations primaire et secondaire . . . 65
4.3.2 Validation des structurations primaire et secondaire . . . 67
4.4 Segmentation en activités . . . 69
4.4.1 Descripteurs audio-vidéo . . . 69
4.4.2 Système de segmentation en activités . . . 71

**Chapitre 5**
**Segmentation autour de la similarité** **75**

5.1 Similarité des documents . . . 76
5.1.1 L'Intersection Quadratique Récursive . . . 76
5.1.2 Application des matrices de similarité aux contenus sonores . . . 78
5.2 Chapitrage . . . 80
5.2.1 Méthode de chapitrage . . . 80
5.2.2 Résultats de « chapitrage » . . . 83
5.3 Organisation de contenus audiovisuels . . . 85
5.3.1 Système d'organisation de contenus audiovisuels . . . 86
5.3.2 Interface du système d'organisation de contenus audiovisuels . . . 90

**Discussion** **93**

**Conclusion et perspectives** **99**

**Bibliographie** **107**

# Introduction générale

## 1 Contexte général

~~Compte tenu de l'accroissement gigantesque du volume de données à traiter, la tâche d'indexation devient extrêmement fastidieuse, et l'automatisation semble désormais indispensable.~~

Ne pas commencer par ce type de phrase serait un sacrilège mais débuter par celle-ci n'est pas très original !

Bien qu'elle justifie une partie de mes travaux de recherche et corresponde au contexte général de ce manuscrit, j'ai volontairement barré cette phrase car elle introduit la majorité (trop ?) des travaux de notre domaine (articles, thèses, habilitations, livres, etc.) : elle demeure néanmoins incontournable.

Mes recherches, au cours de ma thèse, avait un double objectif :

- contribuer à l'**analyse sonore** automatique en proposant de nouveaux outils de recherche de composantes primaires robustes,
- effectuer une **structuration automatique de documents audiovisuels** en utilisant ces briques de base.

J'ai l'impression que presque dix ans plus tard ma ligne de conduite est toujours la même ! Bien évidemment, les composantes sonores se sont précisées, les techniques ont évoluées et les applications se sont diversifiées... Il s'avère que mes premiers résultats servent, la plupart du temps, de point de départ aux travaux effectués depuis ma thèse, comme je vais le montrer tout au long de ce manuscrit.

Mes travaux en analyse sonore m'ont conduit à m'interroger plus précisément sur la caractérisation de l'environnement sonore. La communauté scientifique s'accorde sur les composantes de base, que sont la parole et la musique. Mais, la définition de chacune d'elles (propriétés et limites) n'est pas forcément évidente, d'autant qu'elles ne se manifestent pas sur des segments temporels disjoints...

Voici quelques interrogations qui ont émergé ces dernières années en traitement automatique :

- existent-ils des sous-catégories respectives de la parole et de la musique ?
- le message doit-il être audible (ou compréhensible) pour qu'il s'agisse de parole ?
- le RAP (ou le SLAM) sont-ils de la musique ? Du chant ? Les deux ? Une catégorie intermédiaire ?
- les bruits sont-ils tous informatifs ?
- etc.

Je n'envisage pas de répondre à toutes ces questions dans ce manuscrit (certaines étant même assez théoriques ou philosophiques...), mais de fournir des éléments de réflexion sur les unités sonores et audiovisuelles, au travers d'un traitement automatique du signal approprié.

Un certain nombre de traitements automatiques ont donc pour objectif la segmentation du signal. Elle peut aussi bien être réalisé *a priori* (c'est-à-dire sans connaissance du contenu, et éventuellement de la tâche à accomplir...), qu'*a posteriori* où l'identification des unités est alors simultanée, voire requise. La segmentation peut être vue comme le fait de couper un enregistrement en unités stables, c'est-à-dire en recherchant les frontières de début et de fin : la notion de stabilité (homogénéité) est propre à l'objectif de la segmentation visée.

La structuration est une segmentation dont les morceaux sont ordonnés, organisés et l'homogénéité prend alors un sens plus sémantique.

Dans ce document je ne ferais pas de différence entre la segmentation temporelle et la structuration temporelle et je pense d'ailleurs que **tout est segmentation... À condition de la définir correctement !**

Pour s'en convaincre, voici deux exemples. En musique, structurer un morceau correspond le plus souvent à le segmenter en couplets et refrain. En vidéo, structurer un enregistrement revient très souvent à le segmenter en plans, en scènes ou en émissions. Donc la limite entre ces deux termes n'est pas si marquée que cela...

Le titre de ce manuscrit se place dans cette optique, en utilisant uniquement le terme « segmentation » : **Segmentations sonore et audiovisuelle ?** Celle-ci est appliquée à des contenus sonores et audiovisuels. Le point d'interrogation permet de nuancer mon travail sur chaque mot du titre :

- « segmentations », car il sera question de segmentation au sens large, de la segmentation en zones acoustiques stables à la structuration audiovisuelles de haut niveau,
- « sonore », car il ne s'agit pas d'une revue exhaustive des travaux de segmentation sonore mais plus d'une contribution dans le domaine,
- « audiovisuelle », car la segmentation est effectuée en privilégiant la voie audio sur la vidéo.

Cette ponctuation correspond aussi à un petit clin d'œil (rappel) au titre de l'habilitation de Régine André-Obrecht : *Segmentation et parole ?* [André-Obrecht 93] qui supervise mon travail.

## 2 Contexte scientifique et organisation du document

Depuis le début de ma recherche, j'adopte la majeure partie du temps une méthode : trouver des paramètres « bas-niveau » pertinents afin de limiter le plus possible (voire supprimer) toute phase d'apprentissage supervisé. L'apprentissage supervisé en audio comme en vidéo revêt deux principaux défauts :

- la nécessité d'une base de contenus annotés finement, donc manuellement pour l'essentiel. Non seulement, ceci est synonyme de coût mais également de difficultés de recueillement.
- le temps de calcul afférent aux méthodes de classification actuellement employées.

Mes travaux se sont donc focalisés plutôt sur la paramétrisation que sur les méthodes d'apprentissage et de décision. Bien évidemment, des méthodes de classification dites « classiques » sont utilisées et testées au cours de mes recherches mais la plupart du temps il s'agit de se confronter à l'état de l'art et/ou de valider une approche plutôt que d'un apport scientifique.

Dans la suite de ce document, j'utiliserai plutôt le « nous » que le « je » afin de rendre compte du travail collaboratif effectué au sein de l'équipe SAMoVA entre les thésards et mes collègues enseignants chercheurs autour du thème de la segmentation. Les thèses auxquelles j'ai participées étaient naturellement toutes en co-encadrement.

Ce document est une synthèse de travaux conséquents (thèses, stages de Master, projets, etc.) visant à comprendre le cheminement de ma recherche durant la dernière décennie. Pour plus de détails, je renvoie aux manuscrits d'origine dont les références sont indiquées tout au long du texte...

Comme dans bien des documents de cette nature, ce manuscrit est « segmenté » en deux parties : la présentation scientifique de mes recherches dans le cœur du document et une version étendue de mon *curriculum vitae* dans les annexes.

La synthèse de mon activité scientifique se décompose elle-même en deux parties, à l'image du titre de ce manuscrit.

Lors de la première partie, j'aborde mes différents travaux de segmentation sonore. Les principaux apports sont détaillés vis-à-vis des composantes primaires parole, musique et bruit. Il s'agit principalement de recherches autour de descripteurs sonores robustes pouvant caractériser l'environnement sonore d'un enregistrement.

La seconde partie traite de la segmentation audiovisuelle par l'exploitation de paramètres sonores et visuels dits de « bas-niveau » et la combinaison de ceux-ci avec deux objectifs applicatifs : la segmentation autour des intervenants et l'organisation (structure) d'un document audiovisuel. La contribution scientifique se situe essentiellement au niveau du couplage entre l'audio et la vidéo pour une segmentation multimédia.

# Première partie

# Segmentation sonore

# Introduction

L'analyse d'un contenu sonore (que ce soit de la radio, de la télévision ou tout autre enregistrement ayant une composante audio) implique quasi systématiquement une étape de prétraitement dont l'objectif est l'extraction des zones de Parole, Musique et Bruit (PMB). Grâce à ce travail, des traitements spécialisés au type de contenu peuvent alors être réalisés, par exemple une reconnaissance du locuteur sur les zones de parole et une reconnaissance d'instruments sur les zones de musique...

Les deux premières catégories « parole » et « musique » sont assez bien définies. La parole peut être vue comme une suite d'unités phonétiques pour lesquelles la structure formantique est prédominante. La musique occidentale, quant à elle, possède une structure harmonique avec la note comme unité. La catégorie « bruit » est bien plus difficile à décrire : elle se compose du silence, du bruit de fond, des différents sons environnementaux... Ceux-ci étant très hétérogènes, aucune unité ne se dégage, et le bruit est souvent considéré par élimination comme «  tout ce qui n'est ni de la parole ni de la musique ».

Dans un premier temps et afin de mieux comprendre mes travaux et mes orientations en recherche durant ces dix dernières années, je vais commencer par revenir de manière synthétique sur le système de segmentation PMB que j'ai développé lors de ma thèse. Ensuite, je décris les différentes améliorations (affinages) qui ont été réalisées : chaque composante primaire (parole, musique et bruit) ayant donné lieu à des travaux de recherche spécifiques afin d'en préciser le contenu. Ce chapitre se termine par une discussion autour de mes travaux actuels et les perspectives que je me donne à plus ou moins long termes.

## PMB : les origines !

Il ne s'agit pas d'une segmentation PMB au sens traditionnel de l'expression qui consisterait à déclarer une zone de signal comme appartenant à une des trois classes que seraient la parole, la musique et le bruit ; deux problèmes sont étudiés ici, ils correspondent à deux segmentations parole/non-parole et musique/non-musique. L'intérêt de ces deux sous-systèmes est double : permettre la distinction des zones pures des zones superposées d'une part et proposer des paramètres spécifiques à chacune des composantes d'autre part. Ces paramètres peuvent être caractéristiques de la composante étudiée de manière intrinsèque ou discriminants par rapport aux autres composantes.

Ce système [Pinquier 03a] est fondé sur l'extraction de quatre paramètres (cf. figure 1) :

- la **modulation de l'énergie à 4 Hertz**. Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hertz [Houtgast 85]. Ces modulations correspondent au rythme syllabique de l'élocution humaine.
- la **modulation de l'entropie**. Des observations menées sur le signal (ainsi que sur le spectrogramme) font apparaître une structure « ordonnée » lorsque le signal correspond à de la musique harmonique. La présence de parole est alors synonyme de « désordre » et pour mesurer ce « désordre », nous avons proposé un paramètre fondé sur l'entropie du signal [Moddemeijer 89].
- le **nombre de segments par seconde** et la **durée de ces segments**. Ces deux derniers paramètres résultent d'une analyse statistique fine du signal audio dans le domaine temporel. L'hypothèse de départ est que le signal de parole comme de musique est décrit par une suite de zones acoustiques quasi-stationnaires : les phonèmes et les notes. L'algorithme de « Divergence Forward-Backward » (DFB) [André-Obrecht 88] permet d'approcher ce type de segmentation. De par leur production, le nombre (et la durée) de segments par unité de temps s'avèrent très différents en musique et en parole.

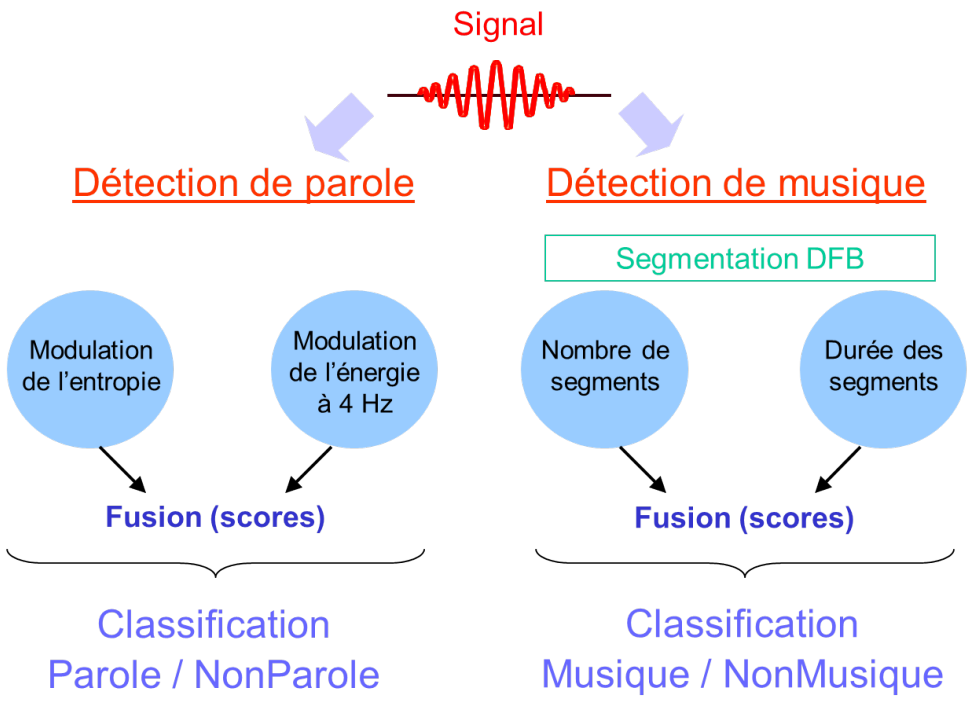



FIGURE 1 – Le système global de fusion de paramètres pour la segmentation PMB.

La classification s'effectue sur une zone (fenêtre de décision) d'une seconde avec un décalage d'une trame d'analyse (20 ms). Un simple seuillage sur chaque paramètre permet de conclure en l'absence ou la présence de chaque composante Parole et Musique sur chaque zone.

Une zone étant classée ni parole, ni en musique sera considérée comme du « bruit » ou un son environnemental comme Gygi le propose [Gygi 07] : *environmental sounds is defined as all naturally occurring sounds other than speech and music.*

Ce système est le résultat de différents travaux contrastifs. Des comparaisons à des approches classiques ont été effectuées. Les paramétrisations spectrale et cepstrale (Mel Frequency Cepstral Coefficients, MFCC) couplées à des Modèles de Mélanges de lois Gaussiennes (GMM) ont été évaluées [Pinquier 03b]. Le stage de DEA de José Arias a également permis de tester les Machines à Vecteurs de Support (SVM) [Arias 04]. De plus, différentes méthodes de fusion ont été testées dans le cadre du DEA de Julie Mauclair [Mauclair 03] : principalement la théorie des probabilités et la théorie de l'évidence. Finalement, l'amélioration des résultats n'étant pas significative (par rapport aux coûts de traitement ou d'apprentissage ajoutés), le système effectue simplement une maximisation des scores pour chaque couple de paramètres.

Afin de donner plus de détails sur le système PMB (pour ceux qui n'auraient pas encore lu le Tome 1 de mes travaux de recherche !), je vous renvoie vers ma thèse [Pinquier 04] qui précise les choix effectués et les résultats obtenus.

L'une des particularités de ce système est sans nul doute sa généricité à traiter tout type de contenus sonores. En effet, sans connaissance du type de l'enregistrement et sans apprentissage ou adaptation aux données, les résultats restent néanmoins corrects. Par exemple, lors de la campagne d'évaluation ESTER [1] [Galliano 05], ce système a obtenu les meilleurs résultats sur la tâche Segmentation en Événements Sonores (SES, sachant que les événements sonores étaient la parole et la musique) sans pour autant utiliser le corpus d'apprentissage fourni (contrairement aux autres méthodes).

Cette propriété explique pourquoi ce système sera utilisé de manière importante par la suite : les résultats restent de qualité, quelque soit le signal audio traité.

Ayant remis à plat, les bases de mon travail, nous allons entrer dans le vif du sujet de ce document : c'est-à-dire mes recherches depuis la fin de ma thèse, et en particulier une analyse plus fine des composantes « parole » et « musique ». Nous verrons également des segmentations qui vont au delà de ces composantes primaires.

1. Évaluation des Systèmes de Transcription Enrichie d'émissions Radiophoniques, http://www.afcp-parole.org/camp_eval_systemes_transcription/

# Chapitre 1

# Segmentation de la composante parole

**Sommaire**


**1.1 Segmentation en locuteurs . . . . . 11**
1.1.1 Système SRL initial . . . . . 12
1.1.2 Système SRL enrichi . . . . . 12
**1.2 Détection de zones d'interaction . . . . . 13**
1.2.1 Contexte et définitions d'une unité d'interaction . . . . . 13
1.2.2 Système de détection et de caractérisation des Z.I. . . . . 14
**1.3 Reconnaissance du rôle . . . . . 17**
1.3.1 Définition des rôles . . . . . 17
1.3.2 Paramètres « bas-niveau » . . . . . 18
1.3.3 Système de reconnaissance des rôles . . . . . 19


Mon exploration de la composante parole m'a conduit à trois types de travaux que je présente ci-après. Dans un premier temps, je décris mes apports sur un incontournable outil d'indexation en parole : la Segmentation et le Regroupement en Locuteurs (SRL). Je traite principalement de sa complémentarité avec ma segmentation en parole/non-parole (sous-système PMB, figure 1). Ensuite, j'aborde les retombées possibles d'un tel outil pour des segmentations de plus haut niveau. Ainsi, j'introduis les zones d'interaction entre les locuteurs et le rôle des intervenants dans les émissions de radio ou de télévision.

## 1.1 Segmentation en locuteurs

La segmentation en locuteurs ou plutôt la Segmentation et le Regroupement en Locuteurs (SRL) est une étape très importante en indexation du flux de parole. Celle-ci consiste à découper le flux sonore en segments homogènes les plus longs possibles, homogène au sens où une seule personne parle durant ce segment. Ces segments sont ensuite comparés et rassemblés en une seule classe s'ils appartiennent à un même locuteur.

La SRL est souvent primordiale pour de nombreux autres traitements : transcription de la parole, reconnaissance du locuteurs, etc. Plusieurs applications qui en découlent sont présentées dans la suite du document : l'interaction entre les locuteurs (section 1.2), le rôle des locuteurs

(section 1.3), la détection des intervenants (section 4.1), la caractérisation des intervenants (section 4.2), etc.

Le travail présenté ici correspond à une collaboration avec Elie El Khoury et son encadrante Christine Sénac lors de la thèse de celui-ci [El Khoury 10].

### 1.1.1 Système SRL initial

Dans le système de base de segmentation en locuteurs d'Elie El Khoury [El Khoury 07], après une détection d'activité vocale assez classique par des MFCC et des GMM, le signal acoustique est segmenté avec un double critère GLR-BIC (pour plus de détails sur le Generalized Likelihood Ratio, voir [Gish 91] et sur le Bayesian Information Criterion, voir [Chen 98]). Le regroupement repose également sur le BIC : un regroupement hiérarchique local est suivi d'une regroupement global. La figure 1.1 illustre ce processus de segmentation et regroupement.

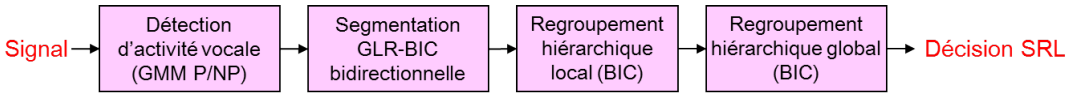



FIGURE 1.1 – Le système SRL initial.

Le système donne de bons résultats sur les zones de parole préparée (type « news »du corpus ESTER [Galliano 05]), mais en présence de parole conversationnelle les scores chutent indubitablement (type « débats » du corpus EPAC [2] [Estève 10]).

### 1.1.2 Système SRL enrichi

Le système SRL intial s'est enrichi de trois manières (cf. figure 1.2).

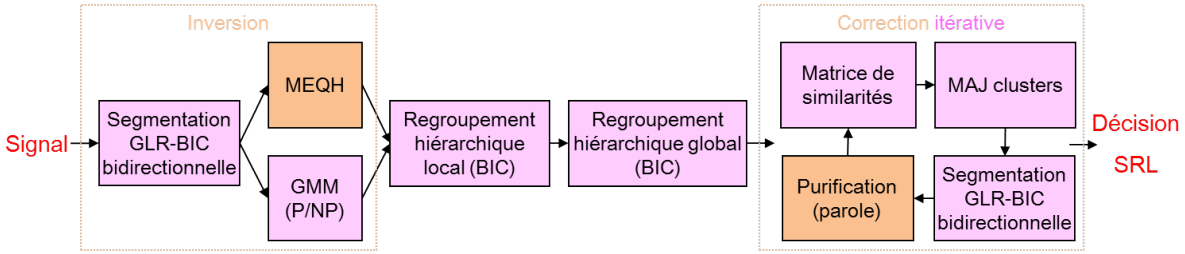



FIGURE 1.2 – Le système SRL enrichi : 1) introduction de la modulation de l'énergie à 4 Hertz, 2) inversion des étapes de segmentation et 3) correction itérative avec purification.

Afin d'une part de robustifier la détection d'activité vocale et d'autre part de mieux gérer les zones de débat (dites « à problème » car contenant de la parole superposée), des paramètres du système PMB présenté en introduction sont intégrés. La Modulation de l'Énergie à Quatre Hertz (MEQH) est couplée au détecteur d'activité vocale classique du système de base (MFCC-GMM).

2. Exploration de masse de documents audio pour l'extraction et le traitement de la PArole Conversationnelle, http://projet-epac.univ-lemans.fr/

Afin de diminuer les erreurs dues à de la parole superposée, notamment en présence de parole conversationnelle, les étapes de segmentation en locuteurs et de détection d'activité vocale ont été permutées. Ceci permet d'avoir des segments homogènes en entrée de la détection de parole plutôt que traiter des trames d'analyse en « aveugle ». Cet échange permet un gain de 1 % sur le score du Diarization Error Rate (DER) aussi bien sur le corpus ESTER que sur le corpus EPAC [El Khoury 09].

Enfin, une étape de correction itérative a également été ajoutée. Celle-ci se décompose en quatre parties :

1. une création d'une matrice de similarité entre les segments et les clusters,
2. une mise à jour des segments à l'intérieur des clusters,
3. une nouvelle segmentation GLR-BIC, avec une évolution de la valeur du paramètre $\lambda$,
4. si la segmentation n'est pas stable, une purification en parole des clusters est effectuée (en contrôlant la détection d'activité vocale à l'aide de la MEQH) et nous recommençons à partir de l'étape 1.

Cette dernière amélioration permet de diminuer le DER de plus de 8 points pour atteindre 16,7 % sur le corpus EPAC. Nous verrons par la suite (section 3.1.2) que la prise en compte des zones de non-parole (musique notamment) dans une approche mixte PMB-SRL peut être bénéfique pour les deux segmentations.

Afin de valider, s'il en est besoin, l'intérêt de la SRL, je vais maintenant décrire deux études qui prennent appui directement sur celle-ci : la détection de zones d'interaction et la reconnaissance de rôles.

## 1.2 Détection de zones d'interaction

Ce travail a été réalisé durant la thèse de Benjamin Bigot [Bigot 11], que j'ai co-encadré avec Isabelle Ferrané, sous la direction de Régine André-Obrecht.

Outre un intérêt d'un point de vue « structuration » (cf. section 4.3), la détection des zones d'interaction aide principalement à la localisation des zones de parole spontanée, en opposition à celles de parole préparée. La parole spontanée [Luzzati 04] correspond à « un énoncé perçu et conçu au fil de son énonciation ». D'ailleurs, comme le rapporte [Dufour 09], un lien existe entre le degré de spontanéité du langage et les performances de reconnaissance. Il conclut 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.

### 1.2.1 Contexte et définitions d'une unité d'interaction

Une **Zone d'Interaction orale** (Z.I.) est une zone temporelle d'un document détectée comme de la parole, durant laquelle seuls deux locuteurs interviennent, et ce de manière alternative. Nous nous affranchissons d'un *a priori* sur le contenu linguistique : peu importe le

message, l'information que nous utilisons indique seulement si un locuteur parle ou ne parle pas à un instant donné. Dans les documents audiovisuels, les interventions des locuteurs peuvent correspondre à des *dialogues* ou à des *monologues*.

Un **dialogue** est 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 à un débat. Aucun autre locuteur n'intervient durant cette unité de dialogue.

Un **monologue** est un discours adressé à une audience, qui ne permet pas l'alternance. La lecture des titres par un présentateur de bulletin d'information est un monologue d'après cette définition, le message étant adressé aux spectateurs.

L'**Unité d'Interaction** (U.I.) est une série de 3 segments de parole formant une alternance entre deux locuteurs différents $loc_j$ et $loc_{j'}$, $(j \neq j')$. Elle est définie par :

$$\{s(i, loc_j) \; - \; s(k, loc_{j'}) \; - \; s(i+1, loc_j)\} \tag{1.1}$$

où $s(i, loc_x)$ est le segment correspondant à la $i^{\text{ème}}$ intervention du locuteur $x$, sous la condition que ces segments de parole ne soient pas séparés par un autre locuteur ou par une zone de non-parole supérieure à une seconde. L'idéal est d'avoir des segments proches les uns aux autres.

Une Z.I. est une suite ordonnée d'U.I. superposées ne faisant intervenir que deux mêmes locuteurs.

### 1.2.2 Système de détection et de caractérisation des Z.I.

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.

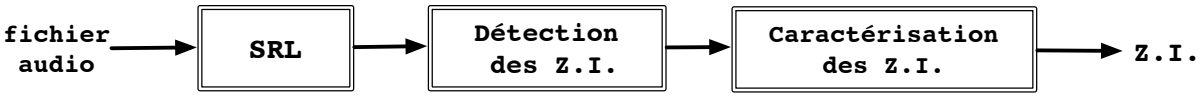



FIGURE 1.3 – Système de détection et de caractérisation des zones d'interaction.

L'approche proposée repose sur trois étapes (voir figure 1.3) :
- une Segmentation et un Regroupement en Locuteurs (SRL, cf. section 1.1),
- une localisation des zones d'interaction associée à une mesure nommée le « niveau d'interactivité », indiquant la longueur de la séquence d'U.I.,
- une caractérisation des zones d'interaction grâce à des informations liées aux locuteurs.

Le niveau d'interactivité ainsi défini permet 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.

La figure 1.4 est un exemple de représentation des zones d'interaction, repérées par leur position dans le document ainsi que leur niveau d'interactivité. Certains locuteurs peuvent être impliqués dans des interactions avec plusieurs locuteurs différents (le *loc*1 dans notre exemple), c'est pourquoi nous avons cherché à caractériser un locuteur en s'appuyant sur « l'ensemble des zones d'interaction ».

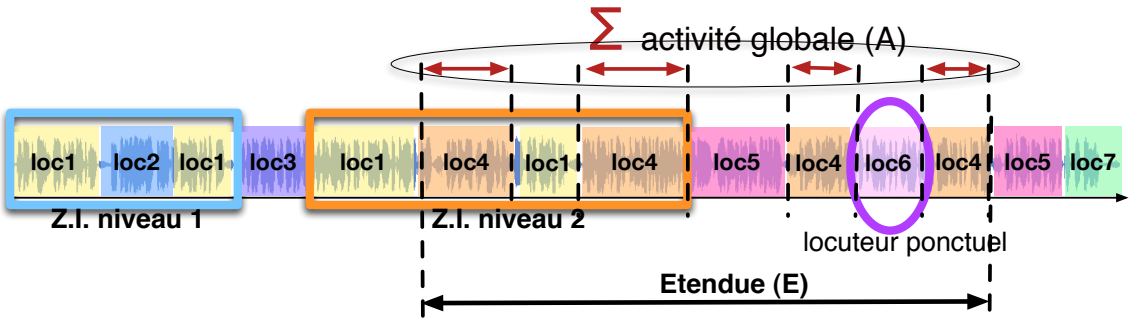



FIGURE 1.4 – Présentation des descripteurs permettant de caractériser les Z.I.

Soient un locuteur $loc_j$, $S_{loc_j} = \{s(k, loc_j), k = 1, .., N_j\}$ l'ensemble de ses $N_j$ interventions et $ZI_{loc_j}$ l'ensemble des zones d'interaction où il est présent.

Nous caractérisons ce locuteur $loc_j$ par les descripteurs suivants :

– l'**activité globale** $A$ du locuteur $loc_j$ est son temps de parole total sur le document :

$$A_{loc_j} = \sum_{k=1}^{N_j} l_{(k,loc_j)} \tag{1.2}$$

où $l_{(k,loc_j)}$ est la longueur du segment $s(k, loc_j)$.

– l'**étendue** $E$ du locuteur $loc_j$ mesure sa durée d'apparition :

$$E_{loc_j} = f_{(N_j,loc_j)} - d_{(1,loc_j)} \tag{1.3}$$

avec $f_{(N_j,loc_j)}$ l'instant de fin de $s(N_j, loc_j)$, dernière intervention du $loc_j$ et $d_{(1,loc_j)}$ le début du segment $s(1, loc_j)$, sa première intervention.

– la **contribution** du locuteur mesure la proportion de l'activité globale $A$ incluse dans ses zones d'interaction :

$$C_{loc_j} = \frac{\text{durée}(S_{loc_j} \bigcap ZI_{loc_j})}{A_{loc_j}} \tag{1.4}$$

Notons que pour tout locuteur n'apparaissant que sur un seul segment, $A_{loc_j} = E_{loc_j}$ et $C_{loc_j} = 0$. Nous appelons ces personnes des **locuteurs ponctuels**.

À partir de ces descripteurs, les locuteurs ont été classés suivant différents types (voir la thèse de Benjamin Bigot pour plus de détails [Bigot 11]) :

- le type 1 correspond aux locuteurs qui ont une activité et une étendue importante,
- le type 2 convient à des locuteurs peu actifs mais très étendus,
- les locuteurs de type 3 sont des intervenants peu actifs et peu étendus (catégorie la plus importante),
- les locuteurs de type 4 ont une activité importante et une étendue plus faible,
- le type 5 correspond aux locuteurs ponctuels (actifs que sur un seul segment), dont leur activité est égale à leur étendue.

La figure 1.5 est un exemple de représentation enrichie des zones d'interaction (épisode du débat de société « Le Téléphone Sonne », sur la station de radio France Inter). Chaque segment 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 un niveau égal à 1). Le type des deux locuteurs impliqués dans chaque zone d'interaction est également indiqué.

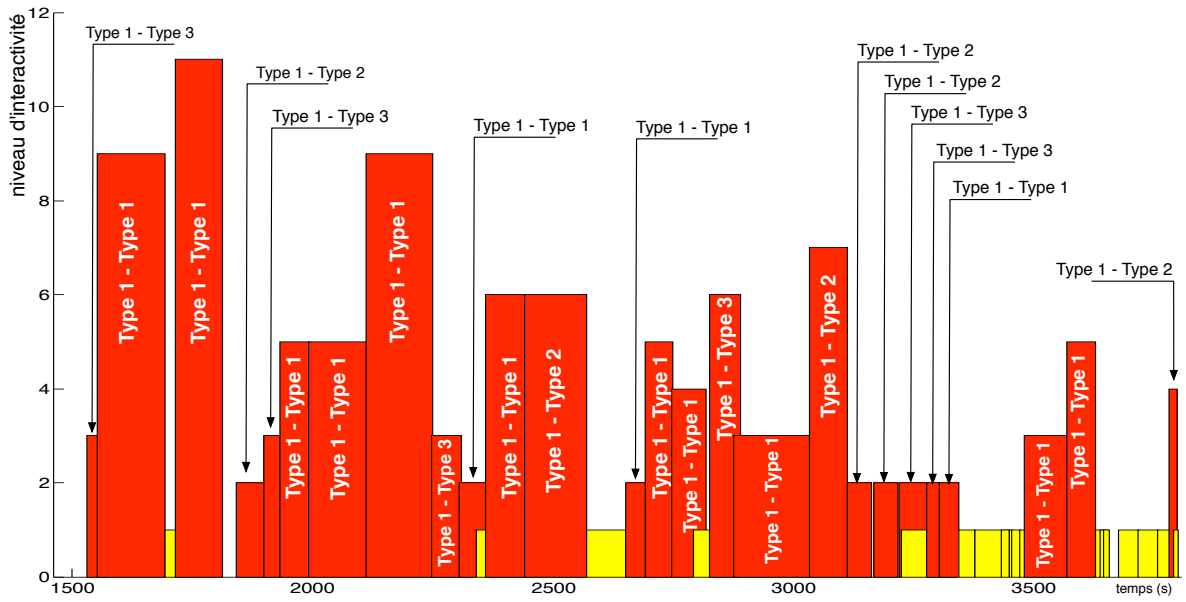



FIGURE 1.5 – Zones d'interaction avec niveau d'interactivité et type des deux locuteurs impliqués, pour le débat de société « Le Téléphone Sonne ».

Ces cinq catégories partagent des points communs avec les rôles des locuteurs dans les documents. En effet :

- le type 1 s'applique bien aux présentateurs, des locuteurs très présents dans un document,

- le type 2 correspond typiquement aux interventions d'un annonceur ou de flash d'informations courts et revenant régulièrement,
- le type 3 coïncide, quant à lui, avec les journalistes qui ont des interventions localisées sur une partie du document,
- le type 4 s'applique à des personnes interviewées,
- le type 5 correspond aux envoyés spéciaux et des journalistes dans les journaux.

Afin d'aller plus loin, c'est donc tout naturellement que je vais enchaîner vers la présentation d'un système de reconnaissance des rôles de locuteurs.

## 1.3 Reconnaissance du rôle

Contrairement aux premières approches de reconnaissance de rôles qui utilisaient une transcription de la parole (tels [Barzilay 00], [Canseco-Rodriguez 04] et [Liu 06]), nous avons choisi de fonder notre travail sur la segmentation en locuteurs, comme Vinciarelli [Vinciarelli 07]. En effet, nous voyons plutôt la reconnaissance de rôles comme un pré-traitement à la transcription, et non le contraire ! Ceci pourrait permettre 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.

Nous faisons l'hypothèse qu'il existe des informations non lexicales sur le rôle des locuteurs, disponibles dans l'analyse de l'interaction, plutôt que dans le contenu du message prononcé par les individus. Par exemple, lorsque nous écoutons 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 !

### 1.3.1 Définition des rôles

Suite à notre travail sur les zones d'interaction (section 1.2), nous avons proposé 5 rôles : présentateur, journaliste ponctuel, journaliste non ponctuel, autre ponctuel, autre non ponctuel. Ces rôles ont l'avantage d'être suffisamment génériques pour correspondre à une majorité d'enregistrements d'une part, et permettre des comparaisons avec les méthodes de la littérature d'autre part.

Un **présentateur** est une personne qui est en charge d'animer (présenter) une émission. Il assure généralement l'introduction et la conclusion de l'émission. De plus, il introduit les autres intervenants et lance les séquences du document.

Un **Journaliste non ponctuel** est typiquement un professionnel, intervenant sur plusieurs tours de parole : un chroniqueur, un intervieweur ou un reporteur.

Un **Journaliste ponctuel** est un cas particulier du journaliste car il n'intervient que sur un seul de tour de parole. Dans cette catégorie de rôle, nous trouvons des envoyés spéciaux, des correspondants, des chroniqueurs (bourse, météo, etc.).

Les deux dernières catégories correspondent à des intervenants qui ne sont ni présentateur, ni journaliste. La classe **autre non ponctuel** correspond 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. Un **autre ponctuel** est un locuteur qui intervient en un seul tour de parole dont la durée peut être très variable, de même que la qualité de l'enregistrement. Cette catégorie correspond souvent à des enregistrements diffusés en différé (extrait de conférence de presse, d'interview, de pièce de théâtre, de film ou d'archives historiques, etc.).

Comme nous venons de le voir, les interventions des locuteurs et leur manière de parler peuvent être typiques de leurs rôles. Nous avons donc extrait un jeu de paramètres « bas niveaux » de différents types (temporels, acoustiques et prosodiques) caractéristiques de leur intervention afin d'analyser chaque locuteur et le classer dans une des cinq catégories de rôle.

### 1.3.2 Paramètres « bas-niveau »

Notre proposition repose sur l'hypothèse que le rôle d'un locuteur reste le même dans un document : un seul rôle est donc attribué à un intervenant du document. Dans ce sens nous fondons notre approche sur l'extraction de trois catégories de paramètres « bas-niveau », calculés individuellement pour chaque locuteur.

#### Paramètres temporels

Ils captent des informations sur la répartition des interventions du locuteur au cours de temps. 14 paramètres temporels sont directement issus de l'organisation temporelle des tours de parole de chaque orateur (SRL). Cet ensemble se compose du nombre de segments $N$, des moyenne/variance/maximum/minimum de ces durées de segments $D_S$ et des moyenne/variance/maximum/minimum des durées d'inter-segments $\Delta_S$. L'activité globale $A$ et l'étendue $E$ sont également calculés (voir section 1.2.2) et trois autres paramètres s'en déduisent :

- le taux d'extinction $T_{ex} = \frac{E-A}{E}$,
- le degré de fragmentation de l'activité $N_S A = \frac{N}{A}$,
- le degré de fragmentation de la segmentation par rapport à l'étendue $N_S E = \frac{N}{E}$.

#### Paramètres acoustiques

Ils caractérisent l'adaptation du locuteur à l'environnement (bruyant ou calme). 10 paramètres acoustiques sont extraits pour chaque locuteur. D'une part, la moyenne et la variance de la puissance du signal sont calculés sur l'intervention complète du locuteur. D'autre part, en nous basant sur les travaux de détection de l'activité vocale de [Atal 76], nous mesurons la contribution du locuteur (zones avec de la parole) et du fond sonore (c'est-à-dire les zones sans parole), en utilisant les statistiques : moyenne/variance/maximum/minimum.

**Paramètres prosodiques**
Ils jugent du professionnalisme du locuteur. Le débit de parole et l'intonation d'un locuteur peuvent varier en fonction de son aptitude à parler en public ou en fonction du niveau de préparation de son intervention. 12 paramètres prosodiques sont utilisés. Le premier sous-ensemble permet de mesurer l'évolution de l'intonation à partir d'une estimation de la fréquence fondamentale ($F0$) du locuteur [de Cheveigné 02] : la moyenne, la variance, le maximum et le taux de zones voisées sont ainsi calculés sur les zones de parole. Le second sous-ensemble 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]. Nous calculons les nombre/moyenne/variance des noyaux vocaliques et des silences, auxquels nous ajoutons les débits de voyelles et de silence.

Au final, nous nous trouvons en présence d'un ensemble de 36 paramètres par locuteur dont nous avons évaluer la pertinence sur différents corpus à travers différentes méthodes d'agrégation : le système ayant été retenu, est un système hiérarchique.

### 1.3.3 Système de reconnaissance des rôles

Il s'agit d'un système hiérarchique (voir figure 1.6) pour lequel toutes les étapes de classification sont ramenées à un simple problème à deux classes, éventuellement précédé d'une phase de réduction de la dimension du vecteur de paramètres.

La première étape, après SRL, consiste à séparer les locuteurs ponctuels des autres. Les locuteurs ponctuels sont divisés en deux catégories : *Journaliste ponctuel vs Autre ponctuel* (Jp *vs* Ap). Les locuteurs non ponctuels, *Présentateur, Journaliste non-ponctuel et Autre non-ponctuel* (P, Jnp et Anp), sont séparés dans un premier temps en deux classes : *présentateur vs non-présentateur*. Puis les non-présentateurs subissent une ultime classification : Jnp *vs* Anp.

À la fin de ce processus hiérarchique, nous nous retrouvons avec les cinq rôles : Journaliste ponctuel, Autre ponctuel, Présentateur, Journaliste non-ponctuel et Autre non-ponctuel.

Nous essayons de tirer profit de chacun des rôles (et de leurs différences !) en utilisant une phase de réduction de la dimension et une phase de classification propres à chaque discrimination de deux classes (rôles).

Les méthodes de réduction de dimension sont de deux types :
- les approches par transformation de paramètres. Nous avons utilisé classiquement une Analyse en Composantes Principales (ACP) [Pearson 01] et une Analyse Factorielle Discriminante (AFD) [Fisher 36].
- les approches par sélection de paramètres. 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.

Les méthodes de classification sont également de deux types : supervisée ou non-supervisée. Dans le cadre de notre travail, trois approches supervisées classiques ont été retenues :

- les Modèles de Mélanges de lois Gaussiennes (GMM) avec un apprentissage par l'algorithme d'Espérance-Maximisation (EM) [Dempster 77],
- les k-plus proches voisins (k-ppv) qui repose uniquement sur l'estimation locale des densités de probabilité [Duda 01] (sans apprentissage),
- les Machines à Vecteurs de Support (SVM) [Vapnik 99] dont le but est de trouver un classifieur qui sépare les données en maximisant la distance entre deux classes. Cette distance s'appelle la « marge » et l'hyperplan séparateur optimal est celui qui maximise la marge.

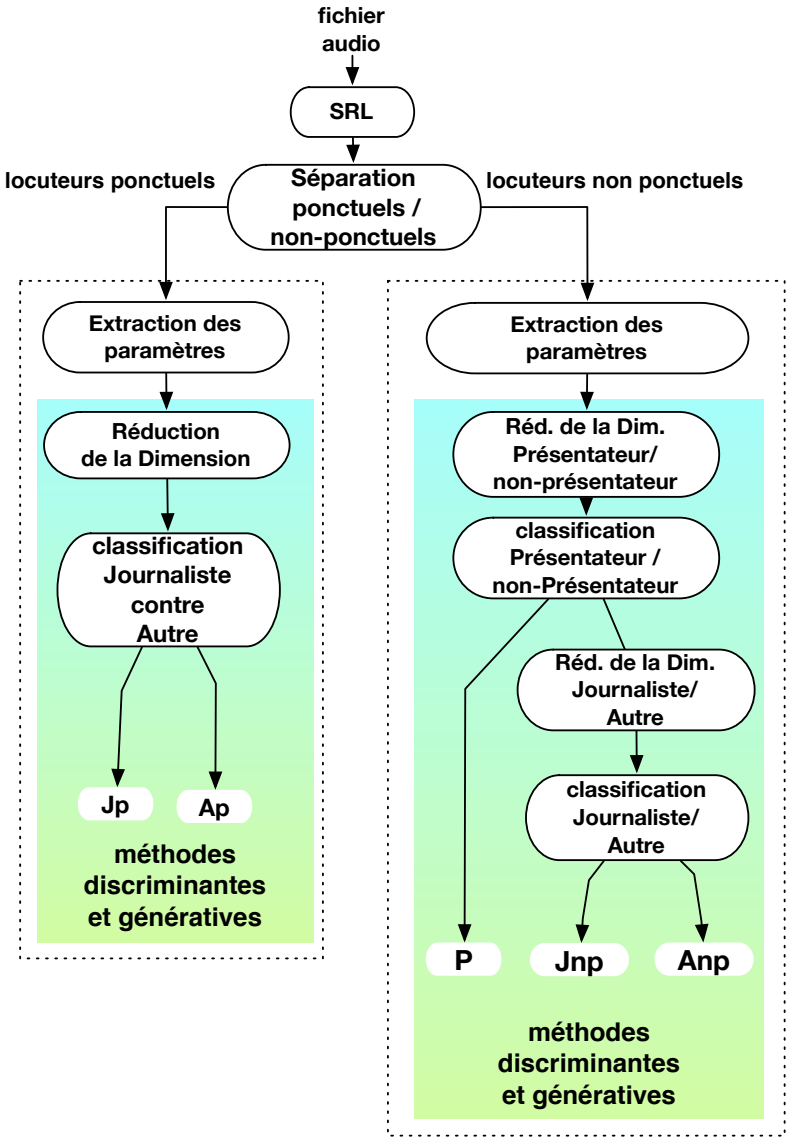



FIGURE 1.6 – Architecture hiérarchique du système de reconnaissance de rôles.

Le meilleur taux de reconnaissance en rôles atteint 81,3 % sur le corpus ESTER2, en utilisant une sélection de paramètres (RSE) couplée à un classifieur SVM linéaire à chacune des étapes. Sur le corpus EPAC, nous obtenons 92 % avec une ACP et un SVM linéaire.

Ces très bons résultats nous permettent d'envisager l'utilisation de ce système de reconnaissance de rôles dans un système de structuration des documents audiovisuels que nous verrons dans la section 4.3.

Pour plus de détails sur ce travail, voir la thèse de Benjamin Bigot [Bigot 11].

# Chapitre 2

# Segmentation de la composante musique

**Sommaire**


**2.1 Discrimination monophonique/polyphonique . . . . . 24**
2.1.1 Système « monopoly » . . . . . 24
2.1.2 Performance . . . . . 26
**2.2 Détection du chant . . . . . 26**
2.2.1 Vibrato . . . . . 27
2.2.2 Système de détection du chant . . . . . 29
**2.3 Détection du chœur à l'unisson . . . . . 30**
2.3.1 Introduction . . . . . 30
2.3.2 Système de détection du chœur à l'unisson . . . . . 30


Les enseignants chercheurs et chercheurs de l'équipe SAMoVA travaillent depuis de nombreuses années sur la composante élémentaire « Parole » du flux sonore et plus précisément sur la composante élémentaire pure (plus de 20 ans pour certains !). L'analyse de la composante élémentaire « Musique »fait son apparition dans l'équipe avec mes travaux sur le segmentation PMB.

L'objectif de départ est double. D'une part, il s'agit d'explorer la composante musique et d'essayer d'en extraire des macro-segments. Il apparait très vite qu'une manière de définir ces macro-segments consiste à déterminer le nombre de sources. Dans un premier temps, nous nous sommes limités à déterminer s'il y a une ou plusieurs sources harmoniques dans la musique. Le problème complet reste un verrou scientifique à ce jour.

D'autre part, les segments contenant du chant représentent une cause non négligeable d'erreurs lors de la segmentation PMB. Il semble alors intéressant de le considérer comme une classe à part entière pour améliorer la segmentation en composantes primaires : parole, musique, chant, bruit...

Je vais me focaliser sur trois travaux. Ces travaux ont permis l'émergence de trois contributions principales.La discrimination monophonique/polyphonique et la détection du chant ont été réalisées dans le cadre de la thèse d'Hélène Lachambre [Lachambre 09]. La détection du chœur à l'unisson fait partie de la thèse de Maxime Lecoz en cours de rédaction.

## 2.1 Discrimination monophonique/polyphonique

Un **son monophonique** est un son produit par une seule source harmonique. C'est soit une note jouée par un instrument de musique, soit une note chantée par un chanteur *a capella*.

Les **sons polyphoniques** regroupent tous les autres sons musicaux, c'est-à-dire tous les sons produits par plusieurs sources harmoniques simultanées. Il y a plusieurs sources harmoniques dès lors que plusieurs instruments (orchestre), plusieurs groupes vocaux *a capella* (chorale) interviennent simultanément. Cela inclut les chanteurs accompagnés et les instruments polyphoniques. Notons qu'un piano peut produire un son monophonique et un son polyphonique !

La segmentation monophonique/polyphonique (communément appelé dans notre équipe *monopoly*) est étudiée sous l'angle de la classification.

Il en résulte deux systèmes, selon que nous considérons deux classes :

- monophonique,
- polyphonique,

ou cinq sous-classes (les deux premières étant monophoniques et les trois suivantes polyphoniques) :

- instrument solo,
- chanteur solo,
- plusieurs instruments,
- plusieurs chanteurs,
- chanteur(s) accompagné(s).

### 2.1.1 Système « monopoly »

Notre étude a été influencée par les travaux de Tsai [Tsai 08], à savoir : *y a-t-il une ou plusieurs fréquences fondamentales ?* Notre culture nous a amenés à traiter ce problème par une approche probabiliste, suivant un schéma classique de Reconnaissance des Formes (voir figure 2.1) : une phase de paramétrisation (extraction des paramètres), suivie d'une phase de reconnaissance (décision).

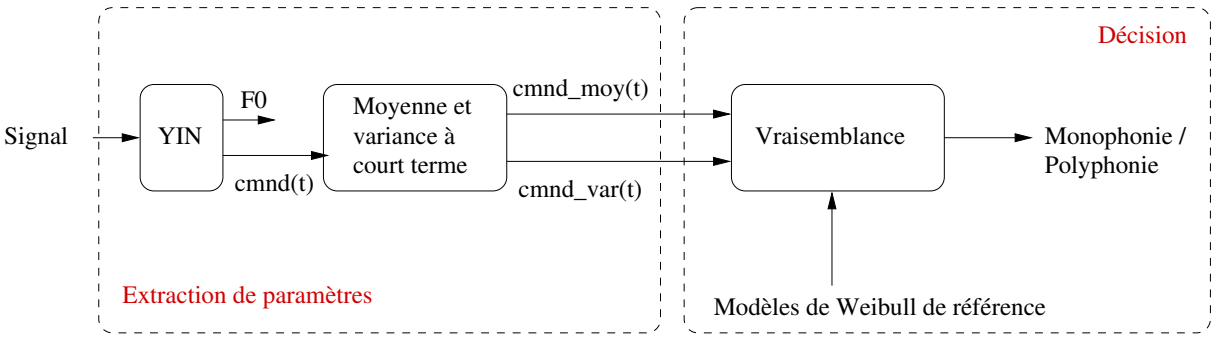
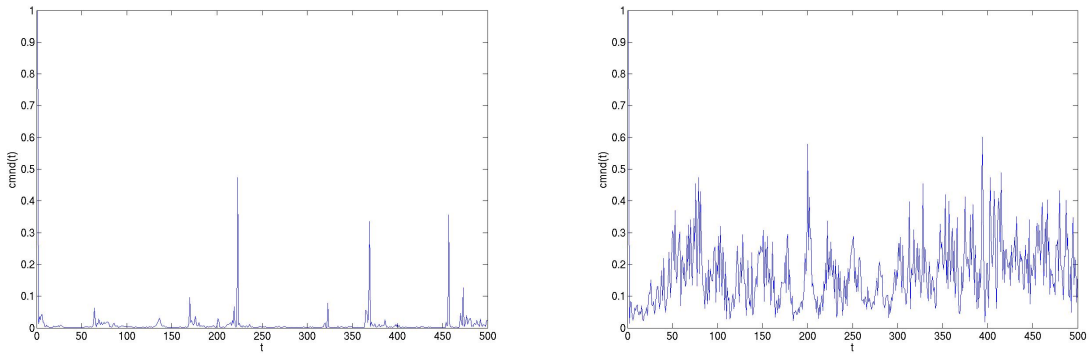



FIGURE 2.1 – Système de discrimination entre sons monophoniques et polyphoniques.

Notre originalité est présente aussi bien dans le vecteur d'observation considéré que dans la modélisation probabiliste choisie. Les paramètres extraits sont la moyenne et la variance à court terme d'un indice de confiance proposé par de Cheveigné [de Cheveigné 02] et les distributions sont des lois de Weibull bivariées. Ce choix de modélisation a été validé théoriquement par le test de Kolmogorov.

**Paramétrisation** Sur chaque trame $t$ de 10 ms, un indice de confiance, noté $cmnd(t)$ est calculé et donne la certitude sur la valeur estimée de la fréquence fondamentale courante ; plus sa valeur est faible, plus la valeur de la fréquence fondamentale est fiable. Sa moyenne et sa variance à court terme, respectivement notées $cmnd_{moy}(t)$ et $cmnd_{var}(t)$, sont calculées toutes les 10 ms, sur une fenêtre glissante centrée sur la trame $t$ de 50 ms, soit 5 trames. Ceci nous donne un vecteur d'observation à deux dimensions $(cmnd_{moy}(t), cmnd_{var}(t))$.

La figure 2.2 illustre notre motivation sur le choix de ces paramètres. Une différence de comportement de l'indice $cmnd(t)$ est notable entre la musique monophonique et la musique polyphonique : dans le cas d'un extrait monophonique, les valeurs sont faibles et varient peu alors que dans le cas d'un extrait polyphonique, elles sont élevées et varient beaucoup.




(a) Musique monophonique (b) Musique polyphonique

FIGURE 2.2 – Variation du $cmnd(t)$ sur des extraits de 5 secondes de signal.

La décision est prise en calculant la vraisemblance des observations sur une seconde, soit 100 couples de paramètres $(cmnd_{moy}(t), cmnd_{var}(t))$, par rapport à chacun des modèles.

**Modélisation** Les histogrammes de répartition des paramètres nous ont amenés à choisir de les modéliser par des lois de Weibull bivariées.

La distribution de Weibull bivariée que nous utilisons a été proposée par Lu [Lu 90]. La fonction de répartition est donnée par :

$$F(x,y) = 1 - \exp\left(-\left[\left(\frac{x}{\theta_1}\right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2}\right)^{\frac{\beta_2}{\delta}}\right]^{\delta}\right) \tag{2.1}$$

Pour décrire une distribution de Weibull bivariée, nous avons besoin de cinq paramètres : ceux d'échelle ($\theta_1$ et $\theta_2$), ceux de forme ($\beta_1$ et $\beta_2$) et celui de corrélation ($\delta$).

Une forte contribution a été de proposer une méthode d'estimation des paramètres de la loi de Weibull bivariée par la méthode des moments [Lachambre 11].

### 2.1.2 Performance

Tout d'abord des comparaisons avec diverses méthodes dites « état de l'art » nous ont permis de valider les différentes étapes de notre méthode :

- la paramétrisation, par une comparaison à des MFCC.
- la modélisation par lois de Weibull, par une comparaison à une approche par lois gaussiennes bivariées.
- la modélisation de Weibull bivariée, par une comparaison à une approche par deux lois de Weibull univariées indépendantes.
- l'approche probabiliste, par une comparaison à des SVM. Nous avons testé trois noyaux : Gaussien, Polynomial et Sigmoïde.

Pour l'approche à deux classes (*monophonique vs polyphonique*), deux lois de Weibull ont été estimées uniquement à partir 50 secondes (monophonique) et 75 secondes (polyphonique) de signal, soit 5000 et 7500 vecteurs ($cmnd_{moy}(t), cmnd_{var}(t)$). Sur un corpus « fait maison », où les classes sont équilibrées en terme de durée, le taux d'erreur obtenu est de 8,5 %.

Pour l'approche à 5 sous-classes précédemment définies, cinq lois de Weibull sont alors estimées avec 25 secondes de signal chacune, soit 2500 vecteurs. Ce découpage permet un gain relatif de 25 % : le taux d'erreur n'est plus alors que de 6,3 % (contre 19,2 % pour les approches classiques fondées sur des paramètres MFCC et des méthodes GMM ou SVM).

## 2.2 Détection du chant

Le Petit Robert définit le chant par *émission des sons musicaux par la voix humaine*. Il est alors assez normal que d'un point de vue « traitement de signal », les caractéristiques du chant se trouvent entre celles de la parole et de la musique. Dans notre système de segmentation PMB, le chant est souvent reconnu comme de la musique, mais il est parfois pris pour de la parole ! C'est l'origine de notre intérêt pour cette composante sonore. Bien évidemment, la détection du chant peut également être une étape de pré-traitement nécessaire pour la reconnaissance du chanteur et la transcription des paroles.

Notre étude nous a fait prendre conscience d'une caractéristique du chant : un chanteur ne chante pas tout le temps, il fait souvent des pauses, principalement dues à des respirations. Nous proposons donc qu'une seconde de signal soit considérée chantée si du chant est perceptible sur quelques trames d'analyse.

La méthode que nous avons développée se base principalement sur la détection du *vibrato*.

### 2.2.1 Vibrato

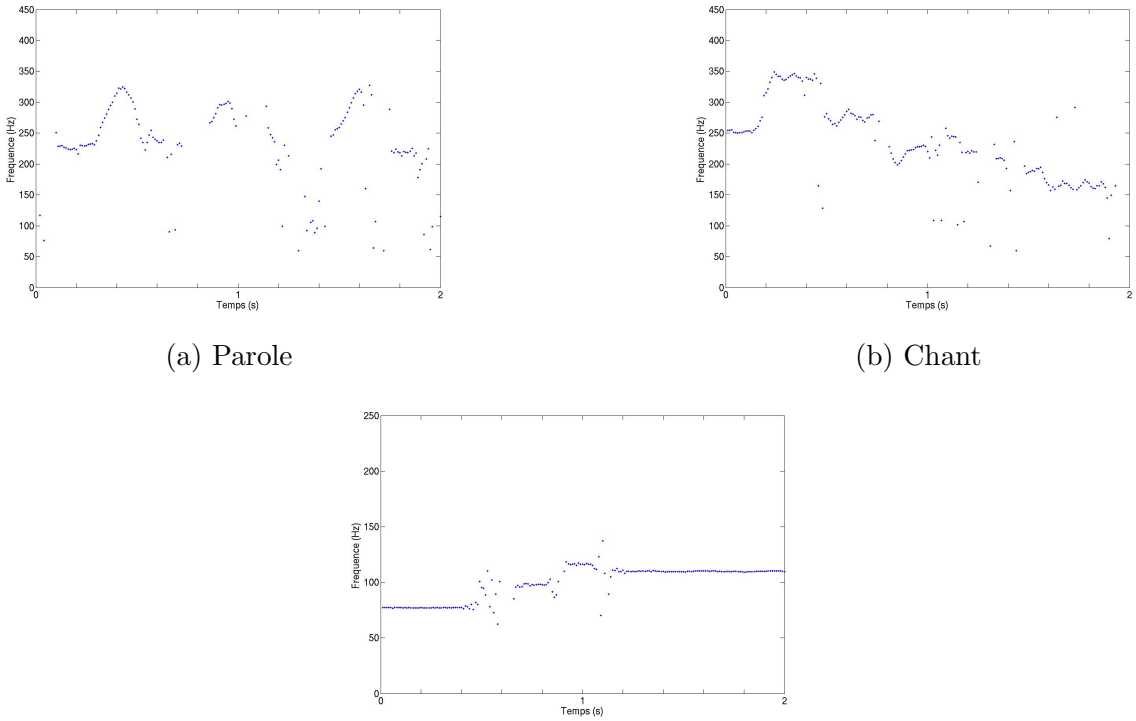
Le vibrato peut être défini comme une oscillation périodique de la fréquence fondamentale d'un instrument ou d'un chanteur [Seashore 38].

Le choix de ce paramètre (le vibrato) s'explique notamment pour deux raisons :

- bien que les chanteurs professionnels puissent le contrôler, il est toujours présent lorsque quelqu'un chante [Timmers 00, Arroabarren 07, Seashore 38],
- la fréquence des oscillations pour le chant est toujours à un rythme compris entre 4 et 8 Hertz, contrairement aux instruments où elle est choisie par le musicien.

Remarque : l'étendue fréquentielle des oscillations est très variable, même pour un chanteur, et peut aller jusqu'à plus d'un demi-ton (140 *cents* [Meron 00]).

La figure 2.3 illustre les différences de comportement de la fréquence fondamentale sur des extraits de 2 secondes de parole (prononcée par une femme), de chant (produit par Barbara) et d'un instrument de musique (contrebasse). La fréquence fondamentale de la parole est instable, celle de la musique est au contraire très constante note par note. La fréquence fondamentale du chant présente clairement du vibrato sur chaque note ; elle oscille périodiquement autour d'une valeur centrale.




(a) Parole

(b) Chant

(c) Musique

FIGURE 2.3 – Fréquence fondamentale d'une personne qui parle (a), d'une personne qui chante (b) et d'un instrument de musique (c). La présence de vibrato est visible pour le chanteur.

Afin de détecter la présence de vibrato, nous suivons la méthode de Gerhard [Gerhard 02] : le vibrato est présent si la transformée de Fourier d'un suivi de fréquence présente un maximum entre 4 et 8 Hertz. La notion de vibrato est intrinsèquement liée à la notion de fréquence fondamentale. Cette dernière s'estimant assez bien en présence d'une source, il est alors possible de définir le vibrato pour des sons monophoniques.

Nous proposons de l'étendre à des sons polyphoniques.Pour ce faire, nous localisons les segments stables fréquentiellement et sur les fréquences saillantes du segment, nous recherchons l'éventuelle présence de vibrato. La localisation de segments stables fréquentiellement repose sur la **segmentation sinusoïdale** introduite par Tanigushi [Taniguchi 05]. Un suivi des principales fréquences est réalisé et permet de définir un segment sinusoïdal au travers de quatre paramètres : un indice de début et un indice de fin correspondant à une localisation temporelle, le vecteur des fréquences et le vecteur des amplitudes résultant du suivi.

Nous en déduisons une **segmentation temporelle** qui consiste à grouper temporellement les segments sinusoïdaux dont les limites (débuts et fins) sont temporellement corrélées. Un segment pseudo-temporel est alors défini par deux limites successives. Il s'en suit deux types de segments :

- les segments longs et stables d'une durée supérieure à 100 ms. Dans le cas d'un son monophonique, ils correspondent à une note ; dans le cas d'un son polyphonique, ils correspondent à un accord, ou à une zone stable harmoniquement (sans changement de note).
- les segments courts. Ils correspondent aux zones de transition, le temps que toutes les harmoniques des notes « sortent » : transition entre deux notes pour un son monophonique, transition entre deux accords pour un son polyphonique.

La figure 2.4 présente un exemple de segmentation sinusoïdale et pseudo-temporelle pour un extrait de 23 secondes de chant monophonique *a capella*.

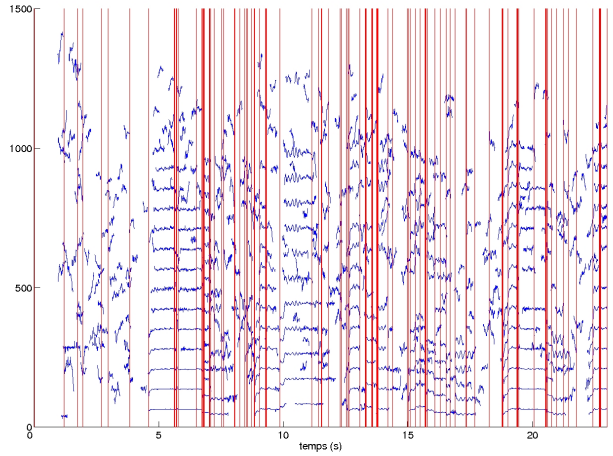



FIGURE 2.4 – Segmentations sinusoïdale (lignes bleues horizontales) et pseudo-temporelle (lignes rouges verticales) d'un extrait de 23 secondes de chant monophonique *a capella*.

Avec la notion de segment pseudo-temporel, nous sommes maintenant en mesure d'élargir le concept de présence de vibrato en introduisant une mesure, le **vibrato étendu**. Ce nouveau paramètre calcule, dans un segment pseudo-temporel, la proportion de segments sinusoïdaux qui ont du vibrato.

Il est calculé de la manière suivante :

$$vibr = \frac{\sum_{s \in \Gamma} l(s)}{\sum_{s \in \Omega} l(s)} \tag{2.2}$$

avec $\Omega$ l'ensemble des segments sinusoïdaux longs (>50 ms) inclus dans le pseudo-segment, $\Gamma$ les segments sinusoïdaux longs avec du vibrato et $l(s)$ la longueur du segment $s$.

### 2.2.2 Système de détection du chant

Le système mis en place (voir figure 2.5) s'appuie sur la segmentation monophonique/polyphonique vue précédemment (cf. section 2.1).

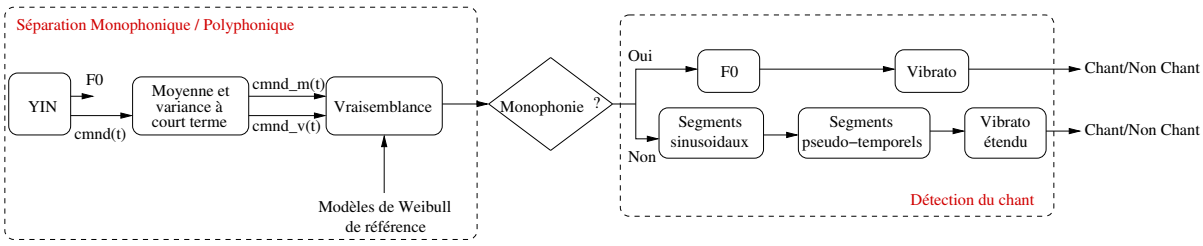



FIGURE 2.5 – Schéma général du système de détection du chant.

Dans le **cas monophonique**, l'estimateur de fréquence fondamentale YIN [de Cheveigné 02] est utilisé. Or, les ruptures brutales dans la courbe de la fréquence fondamentale (dues aux changements de notes) « bruitent » la transformée de Fourier et perturbent la recherche du vibrato (maximum entre 4 et 8 Hertz). De ce fait, nous segmentons temporellement la fréquence fondamentale estimée en « notes », en utilisant une méthode proche des « Note Like Unit » [Ohish 05]. Dès lors, le vibrato est recherché sur chaque note.

Dans le **cas polyphonique**, nous ne pouvons pas nous fonder sur le YIN. Nous cherchons la présence de vibrato sur les segments sinusoïdaux en utilisant le vibrato étendu.

Le système complet obtient un taux d'erreur de 25 %. Sans l'étape préalable de séparation monophonique/polyphonique, le taux d'erreur de la segmentation chant/non-chant était de 29,7 % [Lachambre 07], c'est-à-dire au niveau de l'état de l'art.

## 2.3 Détection du chœur à l'unisson

La musique peut être structurée en trois catégories classiques : les zones de chant, les zones instrumentales et les zones mixtes. Pour préciser cette notion, il peut être intéressant de discriminer les zones monophoniques des zones polyphoniques (cf. section précédente) et bien plus encore de connaître le nombre des chanteurs et/ou des instruments.

Or, une importante difficulté apparaît quand un chœur à l'unisson est observé. Rappelons qu'un **chœur à l'unisson** peut être défini comme un groupe de chanteurs qui chante une même mélodie de manière synchronisée. Dans ce cas, ces chanteurs essayant d'atteindre la même note en même temps, une analyse classique échoue : la zone est jugée monophonique avec un nombre de chanteurs estimé à 1.

Cette section présente comment détecter une telle situation dans un contexte *a capella* (sans instrument).

### 2.3.1 Introduction

Bien que le chœur à l'unisson soit plutôt anecdotique dans la musique occidentale, il est par contre beaucoup plus fréquent dans des contenus ethnomusicologiques, tels ceux présents dans le projet ANR DIADEMS [3] dont nous sommes les porteurs.

Notre approche est fondée sur la détection de divergences dans les harmoniques des différents chanteurs. En effet, bien qu'ils essayent de chanter la même mélodie, de petits décalages temporels ou fréquentiels apparaissent toujours. Ces décalages très fins apparaissent amplifiés sur les harmoniques comme nous pouvons le voir sur la figure 2.6 où leurs suivis sur la zone de chœur comportent des divergences (fractionnements assez visibles au dessus de 1500 Hertz dans cet exemple).

### 2.3.2 Système de détection du chœur à l'unisson

Notre méthode se décompose en trois étapes : une sélection des zones d'intérêt, un suivi de fréquences et une classification.

La **sélection des zones d'intérêt** s'effectue dans le domaine temps-fréquence.

Nous sélectionnons les zones détectées comme monophoniques par la segmentation « monopoly » (voir section 2.1). Une segmentation temporelle est effectuée grâce à l'algorithme DFB déjà utilisé en segmentation PMB (voir section I). Sur chaque zone monophonique, nous nous concentrons sur les 2 segments les plus longs (généralement les phases de « sustain » de note). Sur ces zones la fréquence fondamentale reste stable mais des divergences entre les fréquences des chanteurs apparaissent clairement. Seuls les segments suffisamment harmoniques sont sélectionnés : cette harmonicité est déduite de la moyenne du critère de confiance du YIN [de Cheveigné 02], calculé sur le segment (voir paragraphe 2.2.2 pour plus de détails).

3. Description, Indexation, Accès aux Documents Ethnomusicologiques et Sonores, http://www.irit.fr/recherches/SAMOVA/DIADEMS/

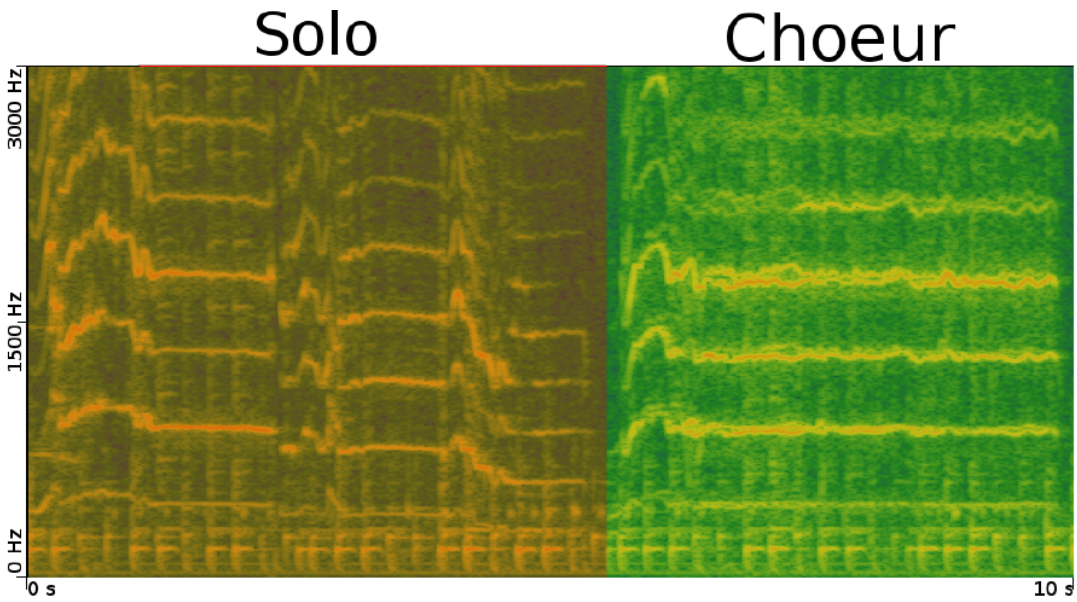



FIGURE 2.6 – Extrait de musique de 10 secondes contenant un chant solo suivi d'un chœur à l'unisson dont le suivi des harmoniques fait apparaître des décalages.

À l'issue de ce traitement, la localisation temporelle de la zone d'interaction est obtenue.

La dernière phase de sélection des zones d'intérêt consiste à localiser les zones fréquentiellement. Pour la ième harmonique de fréquence $i * f_0$, la zone est centrée sur la valeur $i * f_0$ avec une largeur :

$$b_i = min(f_0, i * bw * f_0), \tag{2.3}$$

avec $bw$ un ratio correspondant au pourcentage de $f_0$ à utiliser autour de la valeur de l'harmonique.

Cette augmentation de la largeur de la bande d'analyse en fonction du numéro de l'harmonique permet de garder la même information spectrale dans chaque bande. Cette restriction de la zone d'analyse à la fois sur les plans fréquentiel et temporel permet de concentrer la suite de la détection sur les seules zones pouvant contenir le phénomène d'intérêt. La figure 2.7 illustre cette sélection temps-fréquence des zones d'intérêt (bandes vertes sur le graphique).

La deuxième étape est un **suivi de fréquences** qui se fonde sur la méthode de Taniguchi [Taniguchi 05] (déjà utilisée dans la recherche du vibrato en section 2.2.1). Ce suivi est localisé dans les zones d'intérêts précédemment sélectionnées. Les segments sinusoïdaux suivent l'évolution des principaux pics du spectrogramme et permettent le suivi des zones de fortes amplitudes (voir Figure 2.8). Par la suite, nous considérons qu'un point est un pic principal du spectre, conservé à l'issue du suivi de fréquences.

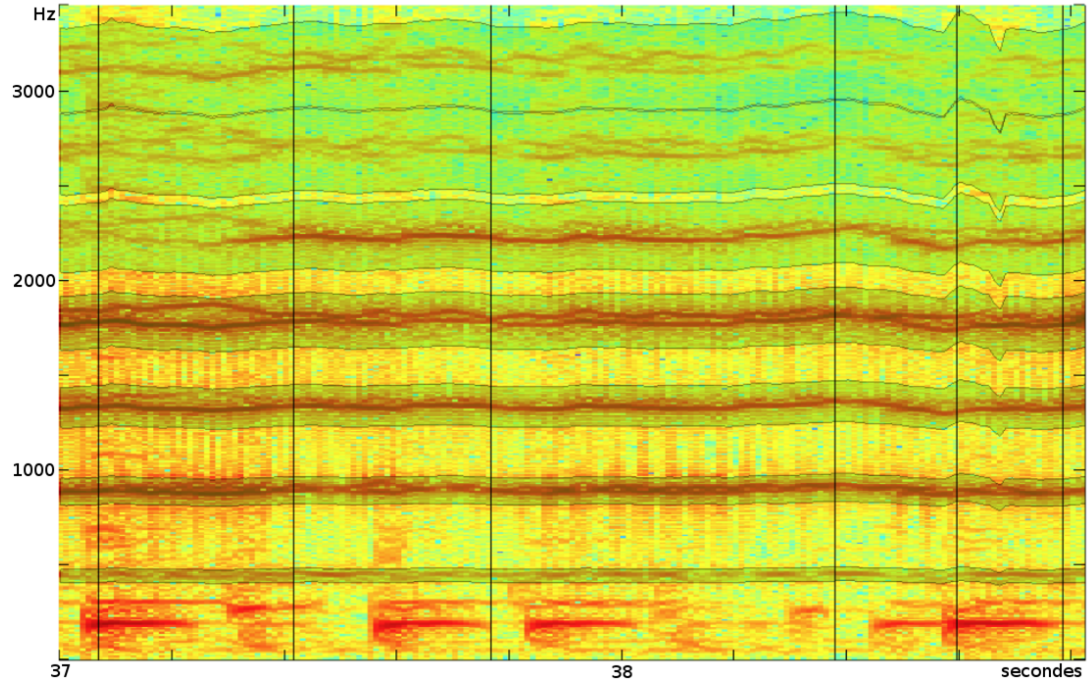



FIGURE 2.7 – Illustration de la sélection des zones d'intérêt : localisations temporelle (traits verticaux noirs) et fréquentielle (bandes horizontales vertes) pour la détection du chœur à l'unisson, sur un extrait de 2 secondes.

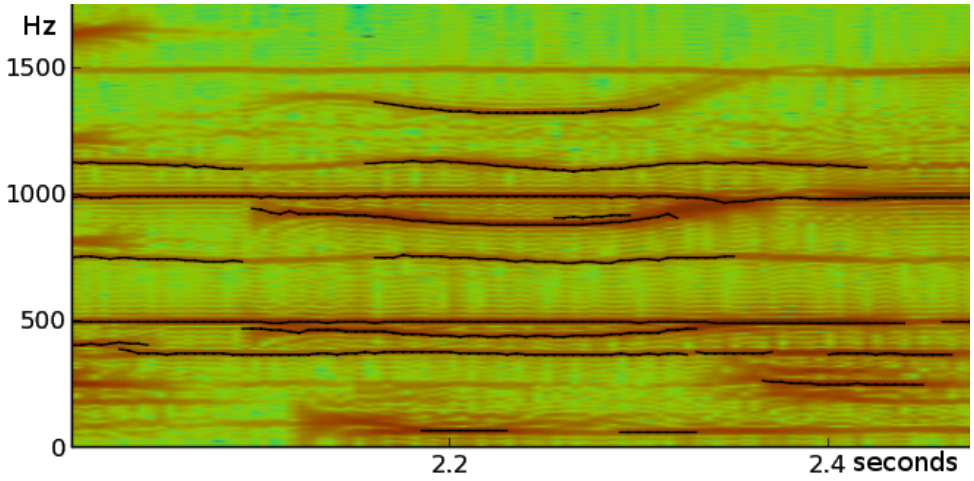



FIGURE 2.8 – Exemple de segments sinusoïdaux superposés au spectrogramme sur une zone temporelle d'une demi-seconde. Les harmoniques de différentes sources sont mises en relief à travers les segments sinusoïdaux estimés (en noir).

La dernière étape est une **classification**. Nous définissons le taux de dédoublement comme le nombre de points présents à chaque instant du spectrogramme (à chaque trame d'analyse) dans chaque zone fréquentielle. Nous calculons ensuite le taux de dédoublement moyen : il s'agit

d'une moyenne fréquentielle (sur l'ensemble des harmoniques) et temporelle (sur l'ensemble de la zone temporelle d'intérêt).

La décision « Solo » ou « Chœur » est prise sur chaque zone temporelle en comparant la valeur du dédoublement moyen à un seuil.

Cette étude est en cours de finalisation mais les premiers résultats sur quelques fichiers sont plutôt encourageants car la segmentation solo/chœur obtient une accuracy de 87 % [Le Coz 12].

La méthode proposée pour la détection de chœur à l'unisson peut être vue comme un raffinage de la décision prise par le système « monopoly ». Dans les zones les plus tenues et aux alentours des harmoniques détectées, nous cherchons à mettre en évidence, grâce à un suivi des fréquences, la présence de segments sinusoïdaux multiples, correspondant à différentes sources. L'étape de classification consiste ensuite à définir si le nombre de divergences détectées est suffisante pour conclure à la présence de plusieurs chanteurs ou qu'au contraire, leur faible présence valide la décision de la classe monophonique.

# Chapitre 3

# Au delà des segmentations primaires

**Sommaire**


**3.1 Segmentation mixte : parole et musique . . . . . 35**
3.1.1 Détection de zones superposées . . . . . 35
3.1.1.1 Introduction . . . . . 36
3.1.1.2 Système de détection de sources multiples . . . . . 37
3.1.2 Fusion PMB/SRL . . . . . 39
**3.2 Segmentation de la composante bruit . . . . . 41**
3.2.1 Détection d'applaudissements et de rires . . . . . 41
3.2.2 Détection de sons d'eau . . . . . 43
3.2.2.1 Système de détection du flot d'eau . . . . . 44
3.2.2.2 Système de détection de gouttes d'eau . . . . . 46


Nous venons de parcourir différents travaux autour des segmentations primaires « parole » et « musique ». Dans ce chapitre, il s'agit d'aller au delà de celles-ci. D'une part à travers leur complémentarité, nous parlerons alors de segmentation « mixte ». D'autre part, nous explorerons les zones qui ne sont ni de la parole, ni de la musique, c'est-à-dire la composante « bruit ».

## 3.1 Segmentation mixte : parole et musique

La première étude est dérivée des travaux que nous avons effectués dans le cadre de la détection solo/chœur, présentée dans la partie précédente : il s'agit d'une méthode permettant de détecter des sources multiples aussi bien en parole qu'en musique. Il s'agit du cœur de la thèse de Maxime Le Coz que j'ai encadré avec Régine André-Obrecht.

La seconde est principalement une fusion de ma segmentation (PMB, section I) et de celle d'Elie El Khoury (SRL, section 1.1) dans le contexte d'une campagne d'évaluation.

### 3.1.1 Détection de zones superposées

L'intérêt de la détection de zones où plusieurs sources harmoniques sont présentes est très important, notamment pour l'amélioration des systèmes de transcription, que ce soit en musique ou en parole. Les sources harmoniques interagissent de manière extrêmement complexe et des stratégies spécifiques doivent être envisagées selon la connaissance du contexte.

En parole, ces zones de paroles superposées sont dites « polluantes » : elles sont difficiles à traiter et malgré leur faible durée elles amènent de nombreuses erreurs de traitement automatique (en SRL, en transcription, etc.). Leur localisation est très importante pour au moins contrôler l'impact sur les zones adjacentes. Une tâche de la campagne d'évaluation ETAPE [4] a d'ailleurs porté sur ce sujet.

En musique, le problème est différent car ces zones peuvent être très longues (orchestre) et informatives : la transcription implique alors des stratégies complexes. La transcription multi-pitch est un sujet très actif de ces dernières années et a donné lieux à plusieurs tâches dans de grandes campagnes d'évaluation, telles MIREX [5] et QUAERO [6].

#### 3.1.1.1 Introduction

Le système que nous présentons vise à localiser temporellement l'existence de plusieurs sources harmoniques par le suivi des fréquences prédominantes dans le signal. Il contribue aussi à attribuer les fréquences prédominantes détectées à chacune des sources présentes. Cette localisation s'effectue sans *a priori* sur le nombre de sources, ni modèle acoustique.

Notre méthode se fonde sur l'analyse en contexte des événements harmoniques à l'aide d'un suivi de fréquences. En effet, les sources sont en général difficiles à discerner dans les zones de recouvrement, pouvoir estimer la façon dont elles se comportent sur les zones contigües permet souvent de lever l'ambiguïté.

La figure 3.1 illustre sur un exemple de parole, la difficulté d'analyse de la zone de superposition. La présence de deux sources différentes est rendu plus évidente par la prolongation des deux sources de part et d'autre du phénomène de superposition (en bleu et en vert).

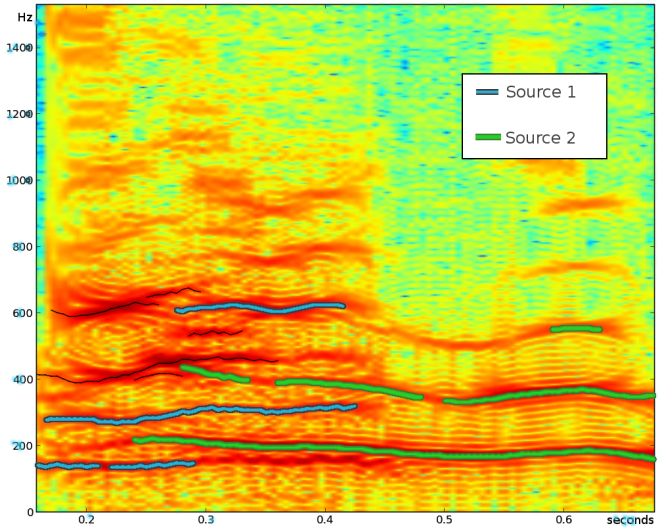



FIGURE 3.1 – Exemple du phénomène de superposition en parole. Le suivi des fréquences avant et après la zone de superposition permet de mettre en évidence l'existence des deux sources.

4. http://www.afcp-parole.org/etape.html
5. http://www.music-ir.org/mirex/
6. http://www.quaero.org/

#### 3.1.1.2 Système de détection de sources multiples

Notre système se décompose en cinq étapes principales : une sélection de zones d'intérêt, une extraction de fréquences candidates, un suivi de fréquences, un rassemblement harmonique et une localisation de superpositions.

La **sélection des zones d'intérêt** consiste à isoler des zones où au moins une source harmonique est présente. Après une segmentation PMB (voir section I), un spectre est alors calculé sur chaque trame de parole et/ou de musique.

Ensuite, trame par trame, nous effectuons une **sélection des pics** spectraux qui possèdent une énergie significative. Nous ne gardons que les pics qui ont une amplitude supérieure à un seuil défini par une fonction linéaire par morceau $th(f)$ dont les paramètres dépendent du pic maximal du spectre analysé $p_{max} = (f_{p_{max}}, a_{p_{max}})$ avec $f_{p_{max}}$ sa fréquence et $a_{p_{max}}$ son amplitude :

$$th(f) = \begin{cases} a_{p_{max}}(\frac{(r_{max}-r_{deb})}{f_{p_{max}}} f + r_{deb}) \textbf{ pour } f \in [0, f_{p_{max}}] \\ a_{p_{max}}(\frac{(r_{fin}-r_{max})}{f_{max}-f_{pmax}} f + r_{deb}) \textbf{ pour } f \in [f_{p_{max}}, f_{max}] \end{cases} \tag{3.1}$$

avec $f_{max}$ la largeur de bande fréquentielle de l'analyse, $r_{deb}$, $r_{max}$ et $r_{fin}$ les ratios respectifs des points de passage $f = 0$, $f = f_{p_{max}}$ et $f = f_{max}$.

Ce seuil sur l'amplitude, illustré sur la figure 3.2, permet de prendre en compte la décroissance d'énergie avec les fréquences et ainsi de faciliter la sélection de pics liés aux sources dans les hautes fréquences.

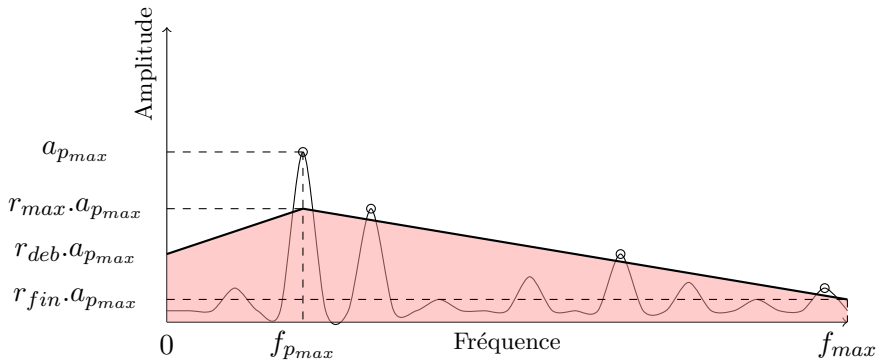



FIGURE 3.2 – Sélection des pics candidats par seuillage dynamique. La fonction linéaire par morceaux utilisée comme seuil est définie à partir des coordonnées du pic principal $p_{max}$.

La troisième étape, le **suivi de fréquences** a déjà été présenté précédemment : il s'agit de la méthode de Tanigushi [Taniguchi 05] (déjà utilisée dans la recherche du vibrato en section 2.2.1 et dans la détection du chœur à l'unisson en section 2.3). Celle-ci permet de relier les pics spectraux des différentes trames en *segments sinusoïdaux* et de quantifier le suivi de l'évolution en temps-fréquences des principales harmoniques présentes sur le segment temporel.

Puis, nous effectuons une **détection de familles harmoniques** : nous cherchons à estimer les segments sinusoïdaux qui sont liés à la même source harmonique. Pour cela, nous utilisons le fait qu'il existe des rapports entiers entre certaines fréquences issues d'une même source. Nous ne pouvons assurer avoir détecté la fréquence fondamentale d'une source et donc seul le rapport de certaines harmoniques est entier.

Ainsi, nous réalisons un graphe où les nœuds sont les segments sinusoïdaux et nous relions uniquement ceux qui sont liés par un rapport fréquentiel entier. Un exemple de graphe représentant les sources pour la musique est présenté sur la figure 3.1.1.2. Les clusters (sources) apparaissent avec des couleurs différentes.

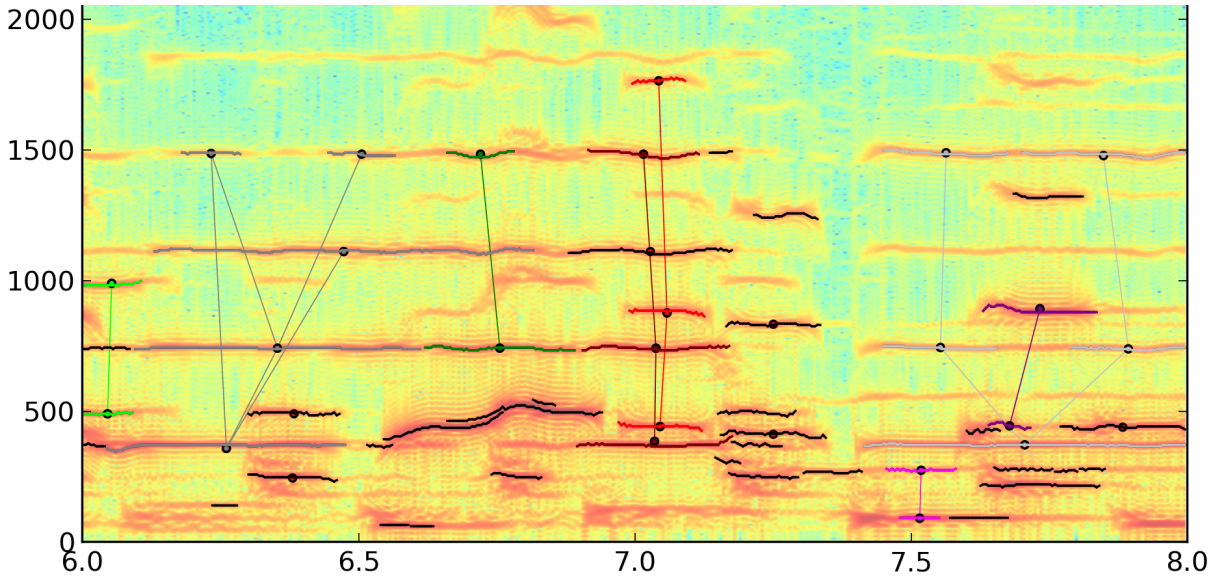



FIGURE 3.3 – Exemple de regroupements harmoniques sur un extrait de 2 secondes de musique polyphonique : 8 clusters sont mis en évidence.

Il s'agit ici d'une contribution forte qui offre des perspectives qui vont bien au-delà de la détection de zones superposées, comme par exemple en vue de détecter du tuilage comme nous le verrons dans les perspectives... Le tuilage peut être défini par une zone de superposition de deux entités $E_1$ et $E_2$, sachant qu'avant la superposition, seule l'entité $E_1$ est présente et qu'après, seule l'entité $E_2$ reste.

La dernière étape consiste à **localiser les zones de superposition**. Jusque ici la méthode était générique : aussi bien applicable à de la parole que de la musique. À partir des différents regroupements établis, la recherche de zones de sources simultanées se fait dorénavant sur un segment temporel qui dépend du type de contenu étudié.

Pour la détection de la parole, concentrée sur de plus petits événements, nous évaluons le nombre de clusters (sources) à chaque instant (trame d'analyse) et s'il est supérieur à 1 nous considérons qu'il s'agit d'une zone de parole superposée.

En ce qui concerne la musique en revanche, la longueur du phénomène nous permet d'utiliser une décision sur chaque seconde. Cette méthode de décision permet alors une plus grande robustesse.

En musique, nous avons évalué ce système sur des corpus très variés : les chansons de l'*Eurovision*, de la musique ethnique issue de la base sonore du CNRS - Musée de l'homme [7]. Les résultats obtenus en musique sont homogènes quelque soit le corpus. Ainsi, le taux d'accuracy (bonne classification) de la détection de musiques superposées est de l'ordre de 70 %.

Obtenir une base d'évaluation de parole superposée est plus difficile compte tenu de la faible durée des segments recherchés. Seule une partie du corpus ETAPE a été évaluée. Les résultats sur les segments de parole sont prometteurs car nous obtenons une accuracy de plus de 78 % en détection de paroles superposées.

### 3.1.2 Fusion PMB/SRL

Alors que traditionnellement et naturellement, la segmentation PMB précède la segmentation en tours de parole, nous avons cherché à combiner les deux segmentations de manière itérative afin de corréler les deux traitements. Le système proposé est une combinaison des deux systèmes précédemment décrits (PMB en section I et SRL en section 1.1). La figure 3.4 illustre cette association de méthodes.

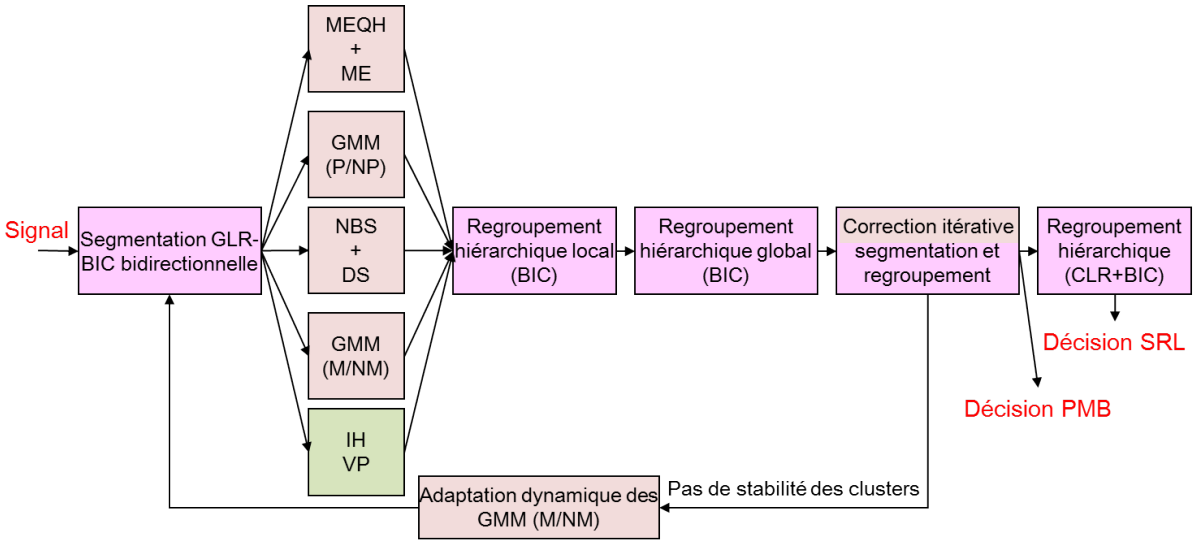



FIGURE 3.4 – Le système mixte PMB-SRL.

Suite à une segmentation bidirectionnelle GLR/BIC, des détections de parole et de musique sont réalisées. Chaque segment est alors annoté en :

- *parole pure* suite à une détection de parole et de non-musique,
- *musique pure* suite à une détection de non-parole et de musique,

7. http://archives.crem-cnrs.fr/

- *parole+musique* suite à des détections de parole et de musique,
- *bruit* suite à des détections de non-parole et non-musique.

L'annotation en parole résulte de la maximisation des scores de vraisemblance de deux sous-systèmes :

- le premier est issu de l'extraction de la Modulation de l'Énergie à Quatre Hertz (MEQH) et de la Modulation de l'Entropie (ME), présentées dans la section I,
- le second est classique, fondé sur des MFCC et des GMM.

L'annotation en musique résulte de la maximisation des scores de vraisemblance de trois sous-systèmes :

- le premier est issu du NomBre de Segments (NBS) et de la Durée des Segments (DS) suite à l'algorithme DFB [André-Obrecht 88], présentés dans la section I,
- le deuxième est classique, fondé sur des MFCC et des GMM,
- le troisième est la combinaison de deux paramètres, l'Indice d'Harmonicité (IH, issu du YIN [de Cheveigné 02]) et la Variance de la Puissance du signal (VP).

Deux regroupements successifs sont effectués sur les segments ainsi annotés : un regroupement hiérarchique local (comme dans le système de base SRL) puis un autre global en utilisant le critère BIC.

Une correction itérative est ensuite effectuée. Si les « clusters » restent identiques entre 2 itérations successives, la procédure s'arrête et les décisions parole/non-parole et musique/non-musique sont réalisées et un dernier regroupement hiérarchique (CLR+BIC) permet d'obtenir une liste des locuteurs. Sinon, à partir d'une adaptation dynamique des GMM de musique par les segments annotés en « musique » jusque-là, le processus recommence et une nouvelle segmentation GLR-BIC est réalisée.

Remarque : il est important de noter que cette phase d'adaptation des modèles de musique améliore bien entendu la segmentation PMB mais également la SRL ! En effet, cette situation se rencontre lorsque des personnes parlent sur un fond musical (exemple : sur les titres des journaux télévisés).

Ce système a été validé lors de la campagne d'évaluation ESTER2 [Galliano 09]. De très bons résultats ont été obtenus pour chacune des tâches (très proches du meilleur système) :

- détection de parole : taux d'erreur de 1,3 % (meilleur système : 1,1 %),
- détection de musique : taux d'erreur de 5,5 % (meilleur système : 5,2 %),
- segmentation et regroupement en locuteurs : DER de 11 % (meilleur système 10,8 %). Notre dernière version obtient 9,85 % [El Khoury 10].

## 3.2 Segmentation de la composante bruit

Le bruit est sans nul doute la catégorie la moins mature au niveau de l'analyse sonore. Ceci est dû à diverses raisons...

D'une part, cette catégorie est assez mal définie : qu'est-ce que le bruit ? Si nous reprenons la définition de Gygi [Gygi 07], il s'agit d'un son environnemental qui n'est ni de la parole, ni de la musique. Or, un son environnemental peut, suivant le contexte, être perçu de manières très différentes... Par exemple, l'effet « cocktail party » peut être soit perçu comme du bruit, soit comme de la parole superposée suivant son niveau. De la même façon, au Mali l'utilisation du pilon sur des oignons peut correspondre à un chant de travail.

D'autre part, le bruit étant très hétérogène, il est alors assez difficile de proposer des traitements automatiques qui puissent s'y appliquer de manière générique... La reconnaissance de sons spécifiques est ainsi privilégiée : nous parlons alors de tâche d'*Audio Event Detection* (AED). Ces tâches qui permettent de détecter une ou plusieurs sources sonores dans un mélange et de leur associer un label, suivant une liste pré-définie, est un problème très compliqué à l'heure actuelle ; les résultats sont en moyenne autour de 30 % de F-mesure.

Une première étude avait été initié durant ma thèse sur les applaudissements et les rires car ces bruits sont assez typiques des divertissements. Ceci nous a permis de segmenter des émissions télévisuelles en spectacles [Pinquier 04]. Le stage de DEA de José Anibal Arias Aguilar, co-encadré avec Jérôme Farinas et sous la direction de Régine André-Obrecht, a fait progressé ce travail, notamment au niveau de la modélisation [Arias 04].

Par la suite, la thèse de Patrice Guyot, co-encadrée avec Régine André-Obrecht, sur la reconnaissance de sons d'eau nous a amené à des réflexions sur le type des sons : continu, discret... Cette thèse sera soutenue le 21 mars 2014.

### 3.2.1 Détection d'applaudissements et de rires

Le choix initial de la détection de ces sons est motivé par le fait qu'ils sont très présents dans des émissions de télévision dites de « plateau » (divertissement, jeu...), et que leur détection révèle la présence d'un événement caractéristique au sein de l'émission telle une performance artistique (musicale, comique, sportive, etc.).

Le signal correspondant aux **applaudissements** est d'un point de vue statistique stable et le contenu spectral est également assez uniforme (cf. figure 3.5).

Par contre, il est difficile visuellement, que ce soit sur le signal ou le spectrogramme, de reconnaître des **rires** (cf. figure 3.5). En effet, le signal est fortement bruité et non-stationnaire : la modélisation de ce son clé apparaît d'ores et déjà difficile. Les rires présentent une grande variabilité naturelle car les personnes rient de plusieurs manières différentes alors que la manière d'applaudir semble universelle.

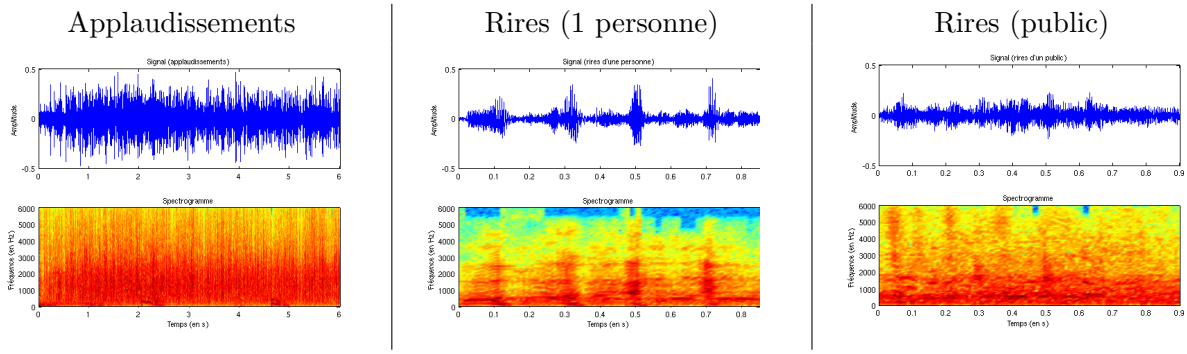



FIGURE 3.5 – Signal et spectrogramme associé d'extraits d'applaudissements (6 secondes) et de rires soit d'une personne, soit d'un public. Les extraits de rires durent environ une seconde.

Le système mis en place, quel que soit le *bruit* étudié, est un système classique de reconnaissance des formes avec une phase paramétrisation et une autre de classification qui permet de segmenter le flux sonore en classe et non-classe (voir figure 3.6). Cette méthode est, **une fois n'est pas coutume**, fondée sur un apprentissage afin de créer les modèles de bruit et non-bruit, en l'occurrence ici applaudissements et non-applaudissements d'une part et rires et non-rires d'autre part. Bien évidemment, une phase d'étiquetage manuel est indispensable à l'apprentissage de chacun des modèles.

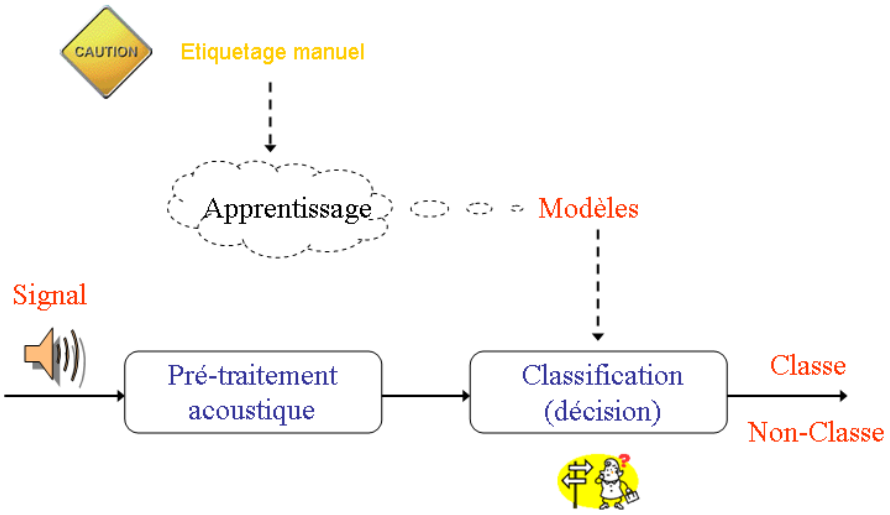



FIGURE 3.6 – Schéma général du système de détection de « bruit ».

Différentes paramétrisations temporelles et fréquentielles ont été testées. Les meilleurs résultats ont été obtenus avec des coefficients spectraux (28 premières valeurs du spectre plus l'énergie) sur des fenêtres d'analyse de 1024 points. La répartition des filtres est linéaire par morceaux et les centres sont : 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, 2300, 2600, 2900, 3200, 3600, 4000, 4400, 4900, 5400, 5900, 6400, 6900 et 7400 Hertz.

Deux types de modélisation ont été étudiées : les Modèles de Mélanges de lois Gaussiennes (GMM) et les Machines à Vecteurs de Support (SVM). Les résultats sont assez comparables entre les 2 approches : les meilleurs scores sont obtenus avec un noyau gaussien pour les SVM et avec 128 gaussiennes (matrice de covariance diagonale) pour les GMM.

La différence se fait principalement sur la quantité de données d'apprentissage utilisée. En effet, avec environ dix fois moins de données d'apprentissage, les SVM obtiennent des scores équivalents. C'est d'autant plus intéressant que l'annotation manuelle est toujours coûteuse...

Globalement, les scores sont excellents avec une F-mesure de 98,5 % pour la segmentation en applaudissements et plus de 97 % pour celle en rires. Ceci nous laisse à penser que notre approche peut être généralisée à d'autres types de bruits. Par contre, si nous analysons plus finement chacun de ces bruits (segment par segment plutôt que de manière temporelle comme avec la F-mesure), les résultats sont moins flatteurs...

En effet, autant pour les applaudissements les résultats restent très bons avec 85 % de segments bien détectés, autant pour les rires seuls 66 % des segments sont correctement retrouvés.

Ceci s'explique assez bien. D'un côté, les applaudissement sont plutôt identiques quel que soit la personne qui frappe dans ses mains et lorsqu'un public applaudit les applaudissements se synchronisent assez naturellement.

D'un autre côté, les rires sont plutôt hétérogènes d'une part entre les personnes et d'autre part entre une personne et un groupe de personnes. Il convient toutefois de noter que les segments les plus significatifs (plus longs) sont bien détectés et identifiés.

Par la suite, nous continuerons à utiliser cette approche pour détecter d'autres bruits (sons d'eau et d'aspirateur notamment) mais plutôt comme un système de référence car cette approche ne nous satisfait pas pleinement. D'une part, l'annotation spécifique pour l'apprentissage rend difficile la généralisation à d'autres types de bruits ou aux changements de conditions d'enregistrements. D'autre part, la paramétrisation « universelle » complique l'explication des phénomènes sonores caractérisés.

Désormais, nous essayons plutôt d'utiliser des approches fondées sur une observation du signal, comme c'est le cas dans la détection des sons d'eau.

### 3.2.2 Détection de sons d'eau

L'une des particularités de ce travail réside dans son cadre applicatif : le projet ANR Blanc IMMED [8]. L'objectif est d'indexer les activités du quotidien d'une personne en vue d'un diagnostic de démence (évaluation des troubles). Ainsi, une vidéo, via un dispositif porté, est réalisée au domicile du patient puis une indexation automatique de ces vidéos en activités est effectuée. Ces vidéos indexées permettent ensuite aux spécialistes de visualiser les patients effectuer des activités dans leur environnement habituel, le disgnostic devient plus intéressant et pertinent.

8. Indexation de données MultiMédia Embarquées pour le Diagnostic et le suivi des traitements des démences http://immed.labri.fr/

Dans ce contexte, de nombreuses tâches quotidiennes ont un rapport avec l'eau : se laver les mains, faire la vaisselle, se brosser les dents, etc.

La reconnaissance de flot d'eau pour des applications médicales a déjà été abordée dans plusieurs études scientifiques. Certaines approches utilisent des capteurs placés directement sur les tuyaux pour détecter l'utilisation de l'eau [Fogarty 06]. D'autres utilisent des microphones placés près du bassin pour reconnaître les activités liées à l'eau [Chen 05]. Dans une étude plus proche de notre application, Taati utilise une caméra placée au-dessus d'un lavabo pour détecter l'activité « se laver les mains » [Taati 10].

Ces précédentes études présentent toutefois le point commun d'avoir été effectuées dans des lieux uniques. Ainsi les données servant à modéliser les sons d'eau et les données visant à tester les systèmes ont été enregistrées dans les mêmes conditions. Celles-ci permettent une utilisation satisfaisante des méthodes d'apprentissage automatique, par exemple des SVM. Dans notre projet, chaque patient étant filmé dans un lieu différent, les données obtenues sont donc très hétérogènes. Aussi nous avons décidé de nous éloigner de notre approche précédente (plutôt « classique ») pour privilégier une approche « bas-niveau » et robuste au changement de domicile.

#### 3.2.2.1 Système de détection du flot d'eau

Dans un premier temps, nous avons testé différents descripteurs acoustiques qui pouvaient caractériser la forme bruitée et continue des sons de flot d'eau. Bien que des descripteurs usuels, tels que le *Zero Crossing Rate* ou le *Spectral Centroïd*, étaient appropriés à la détection de zone de flot d'eau, malheureusement lorsque l'eau est combinée à d'autres sons (telle la parole) les performances s'effondrent (voir figure 3.7).

Le *spectral flatness*, souvent utilisé pour décrire l'aspect bruité d'un son [Johnston 98], est quant à lui trop instable en présence des autres bruits du corpus (choc, frottement sur caméra, etc.).

Nous avons donc introduit un nouveau descripteur, appelé **Spectral Cover** (ou couverture spectrale), dont les caractéristiques permettent de détecter les sons d'eau tout en étant robuste à la voix. Voici la formule :

$$SC = \frac{\sum_i (ampl(w_i) * w_i)^2}{\left(\sum_i (ampl(w_i)\right)^\gamma}, 1 \leq \gamma \leq 2 \tag{3.2}$$

avec $w_i$ les fréquences issues d'une transformée de Fourier et $ampl(w_i)$ les amplitudes associées. Le paramètre $\gamma$ permet quant à lui d'ajuster la prise en compte de l'énergie du signal.

Comme nous le voyons sur la figure 3.7, contrairement aux autre paramètres, un simple seuillage sur la courbe de la couverture spectrale suffit à isoler la partie de l'eau du reste de l'enregistrement.

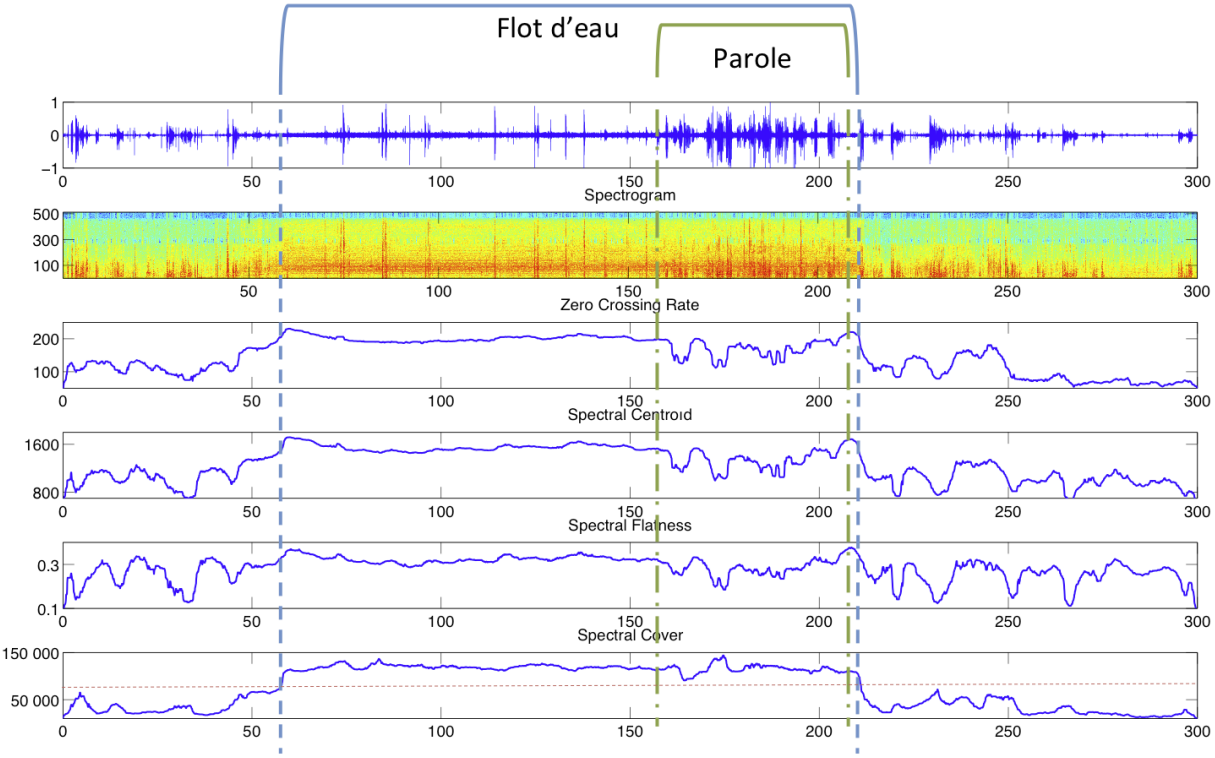



FIGURE 3.7 – Comparaison de la couverture spectrale, « spectral cover », à différents descripteurs acoustiques sur un extrait de 5 minutes.

La couverture spectrale n'est pas uniquement un paramètre pertinent pour la détection de sons d'eau mais elle est également sensible à d'autres types de sons tels ceux des aspirateurs. La figure 3.8 présente le système de segmentation en flot d'eau et en bruit d'aspirateur.

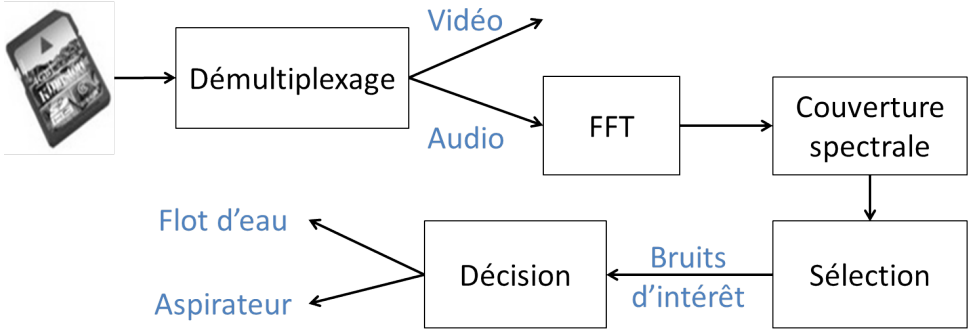



FIGURE 3.8 – Système de segmentation en bruits (sons d'eau et d'aspirateur) d'un enregistrement audio-vidéo.

Dans le cadre du projet IMMED [Mégret 10], la capture audio-vidéo est réalisée par une caméra GoPro. À partir de la carte SD issue de ce dispositif, nous effectuons une étape de démultiplexage afin d'isoler le flux sonore. Suite à une transformée de Fourier, nous calculons notre

paramètre de couverture spectrale. Une phase de sélection (grâce à un premier seuil) nous permet de segmenter l'enregistrement en bruits d'intérêt. Enfin, un second seuil permet de décider s'il s'agit d'un son d'aspirateur ou d'un son d'eau.

La valeur des seuils a été fixée de manière empirique sur un corpus de développement : une vidéo de 40 minutes du projet IMMED.

L'évaluation a été réalisée sur le corpus de test du projet : il s'agit de 20 vidéos d'une durée de plus de 7 heures. Notre système est très performant pour la détection de l'aspirateur avec une F-mesure proche de 95 %. Par contre, la détection de sons d'eau donne des résultats plus faibles : 66 % de F-mesure. Ces résultats sont néanmoins supérieurs aux approches classiques, telle celle utilisée pour la détection des applaudissements et des rires (cf. section précédente) qui au mieux atteint 53 % de F-mesure avec des GMM et des paramètres MFCC [Guyot 12].

La couverture spectrale (par l'intermédiaire de ce système de segmentation en bruits) obtenant des scores satisfaisants, elle fut incorporer dans le système plus large du projet IMMED de reconnaissance d'activités que je présenterai dans la section 4.4.

Après une analyse fine des résultats, il apparaît que la méthode sert essentiellement à détecter le flot d'eau. Or, les activités liées à l'utilisation de l'eau n'impliquent pas forcément un son de flot d'eau : quelquefois il n'y a que des gouttes... Dans la partie suivante, nous allons caractériser plus finement les sons d'eau, en nous basant sur des éléments acoustiques.

#### 3.2.2.2 Système de détection de gouttes d'eau

L'acoustique des sons de liquide, particulièrement de l'eau, a été étudiée depuis de nombreuses années [Bragg 20]. Il peut être étonnant de constater que l'eau en elle-même ne produit pratiquement aucun son. Les sons de liquide viennent ainsi principalement de la vibration de bulles d'air dans l'eau. Ils sont donc constitués d'une multitude d'éléments sonores discrets dont la localisation dans le plan temps-fréquence est reconnue comme étant un indice perceptif permettant de reconnaître les sons de liquides [Geffen 11].

Une bulle d'air apparaît généralement lors de la chute d'une goutte d'eau dans l'eau. Lorsque la bulle d'air monte à la surface, la masse d'eau qui recouvre la bulle diminue, et la fréquence de vibration de la bulle augmente.

Nous pouvons observer ce phénomène dans un son de goutte d'eau, comme sur la figure 3.9 où la montée en fréquence succède à l'impact de la goutte dans l'eau au temps 0.

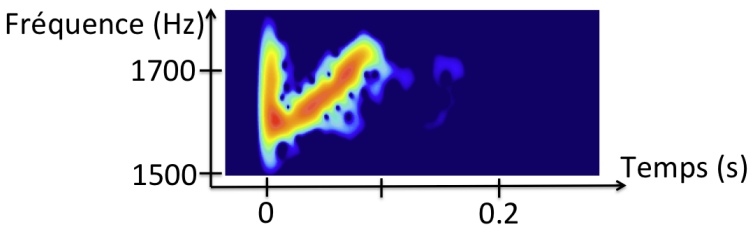



FIGURE 3.9 – Spectrogramme d'un son de goutte d'eau d'une durée de 100 millisecondes.

À partir d'hypothèses provenant des modèles physiques, nous avons proposé un système de détection des sons de bulles d'air. Il se décompose en trois parties : une sélection de candidats dans un banc de filtre fréquentiel, une décision dans un plan temps-fréquence et une étape finale de post-traitement (cf. figure 3.10).

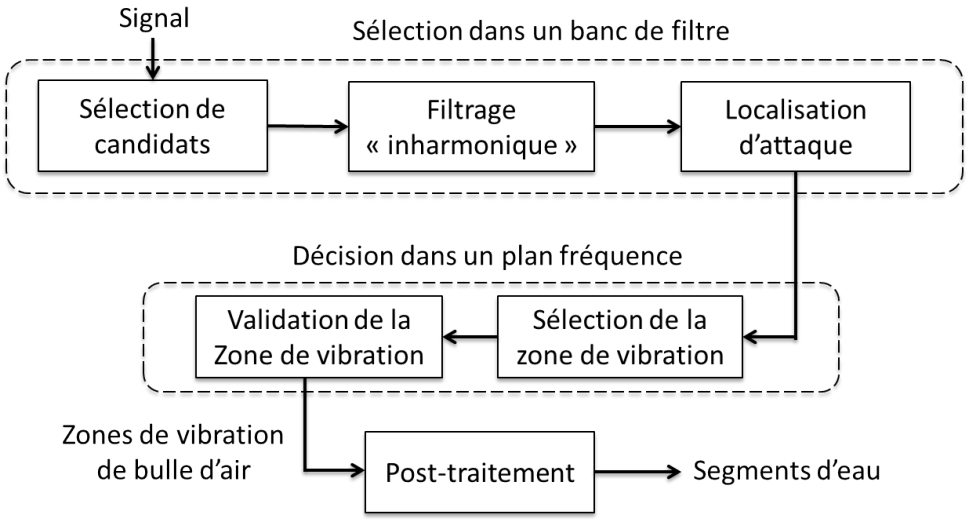



FIGURE 3.10 – Système de détection de gouttes d'eau (bulles d'air).

Nous supposons que toute l'énergie d'une bulle est contenue dans une seule bande de fréquence du spectrogramme à un instant donné. Nous utilisons un banc de filtres fréquentiels (200 Hertz avec recouvrement sur la moitié) pour **sélectionner des candidats**. Ces derniers sont localisés dans les bandes fréquentielles qui contiennent 85 % de l'énergie du signal en cet instant.

La fréquence de résonance des bulles est liée à leur rayon [Minnaert 33] :

$$f = \frac{3}{r} \tag{3.3}$$

Comme les grosses bulles n'apparaissent que rarement dans la nature (sauf quand un très gros objet tombe dans l'eau), nous supprimons les candidats dont la fréquence se situe dans une bande fréquentielle inférieure à 800 Hertz.

Le son de goutte d'eau n'étant pas harmonique, nous enlevons également des candidats à l'aide du critère d'harmonicité du YIN [de Cheveigné 02].

L'étape de sélection se termine par la localisation précise de l'attaque du son, associée au minimum local d'énergie situé entre le candidat et 100 millisecondes avant celui-ci. Nous supposons alors que le candidat trouvé correspond au début du son de bulle.

Dans la phase de **décision**, nous allons considérer l'ensemble du spectrogramme et associer une zone temps-fréquence à chaque bulle. À partir des modèles physiques de vibration, nous

pouvons exprimer l'amortissement de la vibration en fonction de la fréquence de résonance, et ainsi associer des zones temporelles supérieures pour les grosses bulles qui provoquent des sons graves [Leighton 97]. Nous cherchons donc une durée $t$ tel que :

$$\left| a \sin(2\pi f t) \exp^{-dt} \right| < \epsilon \tag{3.4}$$

avec $f$ la fréquence de résonance, $d$ le facteur d'amortissement, $a$ l'amplitude et $t$ la durée. Ceci est vrai quand :

$$t > \frac{\ln(\epsilon/a)}{d} \tag{3.5}$$

En supposant $a = 1$ et en fixant $\epsilon$, il est alors possible de détecter une zone temporelle candidate.

La taille fréquentielle de la zone est fixée à 500 Hertz, ce qui permet de contenir les variations fréquentielles de la bulle. Nous considérons les zones situées avant (pré-zone) et après (post-zone) la bulle d'air (cf. 3.11). Comme les vibrations sont des éléments discrets, nous supposons que l'énergie sera principalement située dans la zone de vibration de la bulle.

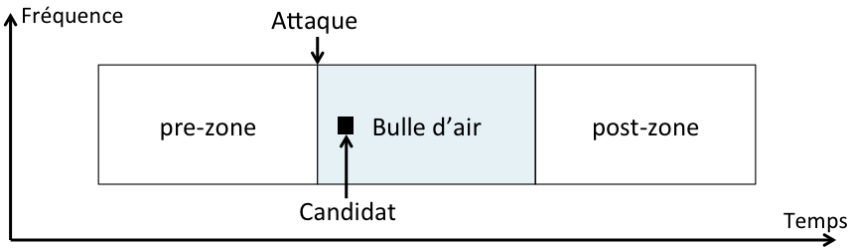



FIGURE 3.11 – Localisation de la bulle dans une zone temps-fréquence.

La dernière étape est un **lissage** afin de supprimer les gouttes isolées.

Sur le corpus du projet IMMED [Mégret 10], nous avons isolé un fichier particulièrement difficile pour le système de détection de flot d'eau présenté précédemment. En effet, sur ce fichier la F-mesure n'est que de 45 %. Le système de détection de gouttes obtient quant à lui un score de 70 %, validant ainsi notre approche [Guyot 13].

Une fusion de ces deux approches continue (flot d'eau) et discrète (goutte d'eau) est en cours d'évaluation.

# Discussion

Dans cette partie, nous venons d'aborder mes différents travaux de recherche en segmentation sonore. À partir du pré-traitement essentiel de discrimination PMB développé lors de ma thèse, chacune des composantes ainsi segmentées a été étudiée.

## Segmentation parole

Sur les zones de parole, j'ai participé à l'amélioration de l'outil classique de segmentation et regroupement en locuteurs. Ce travail a permis de hisser notre système parmi les meilleures technologies françaises en segmentation de parole (PMB et SRL) lors des deux campagnes d'évaluation ESTER. Ce travail de segmentation a ensuite été un fer de lance pour diverses applications : l'interaction entre les personnes a ainsi été étudiée et le rôle des locuteurs a été caractérisé. Ces deux travaux, précurseurs lors du projet ANR EPAC (2007-2010), ont depuis fait des émules dans la communauté scientifique, citons par exemple [Bazillon 11] et [Dufour 11].

Ces segmentations offrent d'autres débouchés :
- segmentation audiovisuelle en combinant le son et la vidéo : nous verrons ceci à travers différentes études dans le chapitre suivant.
- reconnaissance de la parole : aussi bien l'outil PMB que l'outil SRL sont des éléments de base, indispensables à une transcription de bonne qualité.
- etc.

Bien que ces segmentations (PMB et SRL) aient acquises un niveau de maturité satisfaisant depuis près de 20 ans qu'elles sont étudiées, qui leur permet de servir de point de départ à de nombreuses autres applications, la recherche n'est cependant pas terminée pour ces outils « bas-niveau ».

Les contenus traités jusque là étaient plutôt « propres » : bien formatés, souvent composés d'une seule source, dans un environnement non bruité, etc. Dernièrement lors de la campagne d'évaluation ETAPE en 2012, la segmentation de la parole s'est focalisée sur la détection des enregistrement plus complexes, contenant des zones de voix superposées. Nous y avons d'ailleurs participé avec notre système générique de détection de zones superposées 3.1.1. Bien qu'aucun bilan officiel ne soit sorti de cette tâche exploratoire, les résultats sont à l'heure actuelle très faibles.

Il s'agit néanmoins d'un bon début mais il faut aller au-delà.

D'une part, il s'agit d'améliorer les résultats de cette tâche de détection de zones de paroles superposées pour ensuite pouvoir indiquer leur nombre et enfin à long terme les identifier.

D'autre part, il faut se détacher des enregistrements radiophoniques et télévisuels pour proposer des approches plus globales et robustes. Il faut se rapprocher de la « vraie vie » et traiter des enregistrements de réunion, du domicile, de restaurant, etc. Est-ce la séparation de sources qui va résoudre ces problèmes ? Je ne crois pas, en tous cas pas uniquement : une approche plus globale doit être envisagée.

## Segmentation musique

Fort de l'expérience en traitement de la parole de l'équipe SAMoVA à mon arrivée, je fus dans les meilleures conditions pour analyser, proposer et m'enrichir sur cette composante sonore. En revanche, la composante musique n'était pas étudiée jusque là. Quelques travaux dans ma thèse ont été initiés pour la détection de la musique dans la tâche de segmentation PMB. Celle-ci étant très imparfaite sur les zones de chant, nous avons étudié la musique à travers la caractérisation du chant. Ensuite par l'intermédiaire du projet OSEO QUAERO où nous étions évaluateurs de l'ensemble des tâches de musique, l'équipe s'est tournée plus fortement vers ce type de contenu.

Des travaux fondés sur notre expérience en reconnaissance de parole ont permis de proposer un détecteur de tempo [Le Coz 10]. Puis, nous nous sommes penchés sur une problématique forte des systèmes MIR (*Music Information Retrieval*) : la recherche de sources multiples (appelée également *multipitch* pour la détection de plusieurs fréquences fondamentales). Dans un premier temps, une distinction entre les enregistrements monophoniques et polyphoniques a été réalisée. Ensuite, nous nous sommes focalisés sur les zones de musiques superposées avec une détection de chœur à l'unisson, sujet qui à notre connaissance n'avait jamais été traité.

Un parallèle est possible avec la composante parole sur les recherches à venir... Une amélioration des résultats de la tâche de détection de zones de musiques superposées est primordiale afin de pouvoir connaître le nombre de sources (instruments et/ou chanteurs) dans un enregistrement musical. Sachant qu'à plus long terme, l'idéal serait d'identifier chacune des sources.

J'ai initié ce travail par l'encadrement d'un stagiaire de fin d'étude d'ingénieur (François-Xavier Decroix) sur l'analyse et la fusion de différentes méthodes de détection *multipitch*. Nous avons évalué et combiné des approches potentiellement complémentaires afin d'améliorer notre approche actuelle [Le Coz 13]. Nous nous sommes focalisés sur deux approches. La première se fonde sur le calcul d'une fonction de saillance d'une fréquence fondamentale candidate, comme la somme pondérée de ses composantes harmoniques [Klapuri 06] et la seconde utilise différents peignes spectraux [Signol 09] : peignes à dents négatives et peignes à dents manquantes. Les résultats de fusion prouvent la complémentarité des méthodes.

## Segmentation de bruit

La dernière composante sonore sur laquelle j'ai travaillé est sans conteste la plus difficile. En effet, le bruit est une composante très vaste et très hétérogène. Je me suis alors focalisé sur quelques événements sonores spécifiques tels que les applaudissements et les rires qui sont sensés illustrer des zones de divertissement dans des enregistrements radiophoniques et télévisuels. L'approche classique mise en place a prouvé ses limites sur les rires qui sont moins universels que les applaudissements : les rires que ce soit d'une personne à l'autre ou entre une personne et un groupe peuvent être très disparates...

Pour remédier à cela, et en accord avec Pachet qui faisait déjà remarquer en 2007 [Pachet 07], qu'il est préférable d'avoir un paramètre bien choisi plutôt que de multiplier les descripteurs, nous avons confronté différentes variables acoustiques.

Dans notre contexte, ne trouvant pas le paramètre adéquat, nous en avons proposé un : le « spectral cover », qui caractérise les sons dont l'énergie est étendue en fréquences, tels l'aspirateur ou le flux d'eau. La mise en application de ce paramètre nous a amenés à étudier la catégorie « sons liquides », selon la catégorisation perceptive de Gaver [Gaver 93]. Nous avons alors proposé deux systèmes complémentaires, permettant de retrouver des sons d'eau. Le premier capte l'aspect continu du flux d'eau alors que le second détecte le côté discret des gouttes d'eau.

À travers cette dernière étude, il est intéressant de noter que la segmentation automatique du flux audio réalisée par la machine, et la perception auditive humaine doivent être mises en relation afin de proposer des paramètres les plus pertinents possibles par rapport à la tâche ciblée. La prudence est de rigueur, afin de ne pas rechercher ce qui n'existe pas !

## Bilan

Tous mes travaux de recherche ont un point commun : segmenter temporellement un contenu sonore en zones homogènes en vue de fournir une étiquette. Certains sont assez classiques, comme la segmentation PMB ou la SRL, d'autres beaucoup plus originaux, telle la reconnaissance de sons d'eau.

De plus, la majorité des travaux (pour ne pas dire tous...) ne nécessite pas d'apprentissage : ceux-ci sont sensés fournir des systèmes robustes, fonctionnant sur tous types de contenus. Bien évidemment, ceci peut avoir comme inconvénient d'être quelque fois moins performant sur une tâche dédiée, avec un corpus bien ciblé. Mais, être capable de traiter des contenus très volumineux dans des conditions très hétérogènes me semble être bien en adéquation avec notre société actuelle qui produit des quantités de données considérables aussi bien avec des technologies de pointe (dispositifs high-tech) que de simples téléphones portables. Par exemple, ce sont des voyageurs présents sur le quai de la gare de Brétigny-sur-Orge, qui ont filmés la scène de l'accident présentée sur TF1 le 12 juillet 2013 au journal télévisé de 20 heures.

Maintenant que toute personne est potentiellement fournisseur de corpus via les médias sociaux (Facebook, Flickr, YouTube, etc.), la qualité des enregistrements devient assez problématique. La plupart des méthodes de l'état de l'art sont développées dans un cadre d'enregistrements dit « propre » (en studio), et traiter des contenus captés dans la rue, où le bruit environnant n'est pas maîtrisé, devient un challenge très important.

Je suis en train de me confronter à ce genre de problèmes à travers le projet d'ethnomusicologie DIADEMS. Voici quelques exemples de la complexité des contenus qui peuvent atteindre plusieurs dizaines d'années, voire un siècle pour ceux de l'exposition universelle de Paris en 1900 :

- les supports d'enregistrement sont très hétérogènes : analogiques (mécaniques tels le cylindre ou le disque, magnétiques comme les bandes des cartouches ou cassettes) ou numériques (DAT, CD, DVD, etc.),
- des dégradations du contenu telles les parasites apparaissent éventuellement suite à la numérisation des supports analogiques et/ou à leur vieillissement,
- les bruits de l'environnement sont inévitablement présents pour un « enregistrement sur le terrain »,
- notre propre culture « parole »et surtout « musique »doit être remise en cause. Nous qui sommes habitués à traiter de la musique occidentale avec ses règles, nous devons revoir nos référentiels très contraints,
- etc.

C'est ainsi que la segmentation PMB n'apparait plus dans ce contexte comme LA segmentation primaire et qu'il est nécessaire de parler dès lors de zones d'intérêt. Repérer les zones de bruits dits « techniques », tels la saturation, les drops (analogiques ou numériques), les démarrages (et arrêts) de session d'enregistrements... revêt alors une importance primordiale en termes d'usage : la recherche de zones initiales pertinentes !

Ces différents niveaux de segmentation doivent être précisés afin de mieux appréhender la caractérisation de l'environnement sonore d'un enregistrement. Nous évoquerons d'ailleurs quelques pistes possibles dans la fin de ce manuscrit.

# Deuxième partie

# Segmentation audiovisuelle

# Introduction

Après avoir décrit un ensemble de travaux « bas-niveau » fondés sur l'analyse audio dans la première partie, il me semble important de montrer qu'une structuration temporelle ou segmentation d'un niveau supérieur, peut être atteinte à partir de ceux-ci.

D'après Zhang [Zhang 97], l'analyse structurelle d'un document englobe les résultats issus de deux traitements : une segmentation temporelle du document (notion d'unité de base), et une extraction et caractérisation du contenu de ces unités.

Dans cette lignée, voici mes postulats :
- la nature des événements sonores « bas-niveau » est révélatrice d'un schéma de production,
- les enchaînements temporels possibles entre ces différents événements sont aussi révélateurs de la structure,
- l'étude conjointe de l'audio et de la vidéo est incontournable pour atteindre une segmentation adaptée à un document audiovisuel.

Dans ce chapitre, nous allons aborder la structuration audiovisuelle à travers deux aspects.

La première partie est consacrée aux segmentations ayant comme fil conducteur les intervenants. À noter, que nous utilisons le terme « intervenant » pour qualifier une personne qui apparaît à la fois sur la piste sonore en tant que locuteur et sur la vidéo en tant qu'individu physique visible. Le point commun de ces travaux est la présence d'une personne qui apparait soit comme unité de structuration, dans les cas de détection et de caractérisation d'un individu, soit comme la cause de la structuration, dans les cas de communication, d'interaction ou d'activités humaines.

La seconde partie traite de l'organisation (structure) d'un contenu audiovisuel de manière générale, au travers de la notion de similarité. Il s'agit d'une structuration *en aveugle*, sans *a priori* sur le contenu. Le but est de segmenter un enregistrement audiovisuel à travers des unités qui traduisent son agencement, ses enchaînements, sa hiérarchie éventuelle. Nous abordons à travers trois études, la segmentation en programmes et la similarité entre documents.

# Chapitre 4

# Segmentation autour des intervenants

## Sommaire


**4.1 Détection des intervenants . . . . . 58**
4.1.1 Principe . . . . . 58
4.1.2 Matrice de co-occurrences . . . . . 59
**4.2 Caractérisation des intervenants . . . . . 61**
4.2.1 Hypothèses de travail . . . . . 62
4.2.2 Changement d'environnement sonore . . . . . 62
4.2.3 Système de classification IN-OUT-OFF . . . . . 63
**4.3 Structuration en programmes . . . . . 65**
4.3.1 Structurations primaire et secondaire . . . . . 65
4.3.2 Validation des structurations primaire et secondaire . . . . . 67
**4.4 Segmentation en activités . . . . . 69**
4.4.1 Descripteurs audio-vidéo . . . . . 69
4.4.2 Système de segmentation en activités . . . . . 71


La première étape d'une phase de structuration fondée sur la présence d'intervenants, consiste à les détecter. Ce premier travail est issu d'une synergie entre quatre membres de l'équipe SAMoVA avec des rôles assez ciblés :

- Elie El Khoury, en thèse sous l'encadrement de Christine Sénac a amené son savoir faire sur la segmentation en locuteurs [El Khoury 06],
- Gaël Jaffré, au cours de sa thèse sous l'encadrement de Philippe Joly, a été le spécialiste de la détection de personnes (visages et costumes) dans les vidéos [Jaffré 05],
- Frédérick Gianni, comme ingénieur, a développé une interface de navigation et d'annotation fondée sur les intervenants [Gianni 07],
- moi-même, j'étudie la combinaison des informations sonores et visuelles.

Dans une vidéo, il est fréquent que la voix que nous entendons ne corresponde pas à la personne visible à l'écran. Dans son stage de Master, Jérémy Philippeau que je co-encadrais avec Philippe Joly, a proposé une caractérisation des personnes afin de donner une indication sur la présence ou l'absence de celles-ci à l'écran, nous parlons d'intervenants « IN/OUT/OFF » [Philippeau 05].

Ensuite, nous tirons partie du rôle des personnes et de leur interaction afin de structurer le document en deux niveaux de description [Bigot 11]. Il s'agit du travail de thèse de Benjamin Bigot que j'ai co-encadré avec Isabelle Ferrané, sous la direction de Régine André-Obrecht.

Finalement, à travers une collaboration via le projet ANR IMMED précédemment décrit dans la section 3.2.2, nous proposons un système de fusion de différentes modalités sonores et visuelles permettant de structurer une vidéo focalisée sur les activités de la vie quotidienne d'une personne.

## 4.1 Détection des intervenants

Afin de pouvoir détecter les intervenants, nous nous sommes appuyés sur une segmentation en locuteurs côté son et sur une segmentation en costumes côté vidéo afin d'associer automatiquement un visage à une voix.

La segmentation en locuteurs (SRL) a déjà été présentée dans la section 1.1. Voici une description sommaire de la détection de costumes que nous avons utilisée.

La première étape de cette détection consiste à détecter les visages, de manière à trouver les différentes personnes présentes à l'écran. Le logiciel OpenCV [9], fondée sur la détection de visages de Viola et Jones [Viola 01], sert de point de départ.

Le costume de chaque personne est extrait à partir de l'image au prorata de la dimension de son visage. Les paramètres extraits sont les couleurs dominantes calculées à partir d'un histogramme et le spectre d'énergie.

Afin de réduire les fausses détections de visage, les propriétés d'une séquence vidéo sont exploitées en effectuant un filtrage temporel. Dans le même but, deux visages qui possèdent le même costume sont considérés comme associés à la même personne. Pour plus de détails, voir [Jaffré 05].

Notre objectif est ensuite de fusionner, sans connaissance *a priori*, les deux indices produits par les segmentations audio et vidéo, afin de rendre plus robuste l'information portée par chacun.

### 4.1.1 Principe

Les deux index (sonore et visuel) ne sont pas extraits à partir de la même échelle temporelle. Le premier a comme unité le temps : la milliseconde (échantillon) ou plus souvent la centiseconde (trame). L'unité temporelle du second est la cadence (image) ou le plan : l'unité est alors variable ! Afin de les comparer, nous effectuons une normalisation : nous privilégions une échelle commune, à savoir celle de l'image.

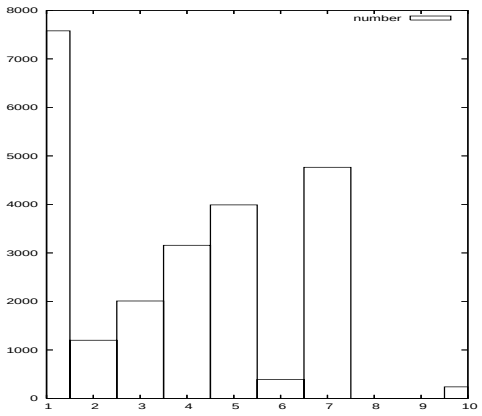
De nombreuses méthodes ont utilisé à la fois des signaux audio et vidéo pour améliorer l'identification des personnes [Albiol 04, Tsekeridou 01, Li 01]. Cependant, les traitements sont effectués quand les deux caractéristiques visuelles et vocales sont présentes, et il est fait l'hypothèse que la voix courante correspond au visage à l'écran. Dans les séquences réelles, cette

9. http://opencv.org/

hypothèse est souvent violée : il est très fréquent de trouver des séquences où les personnes apparaissant ne parlent pas pendant plusieurs images ou plusieurs plans. En outre, il est également habituel que la voix entendue n'appartienne pas à la personne présente à l'écran.

Comme nous pouvons le voir sur les histogrammes de la figure 4.1, les résultats seraient erronés si de manière simpliste, nous attribuions la voix la plus fréquente à l'apparition visuelle la plus fréquente. Ainsi, au lieu de travailler directement sur ces histogrammes, nous proposons de calculer une **matrice de co-occurrences** entre les deux index.

Index audio




Index vidéo

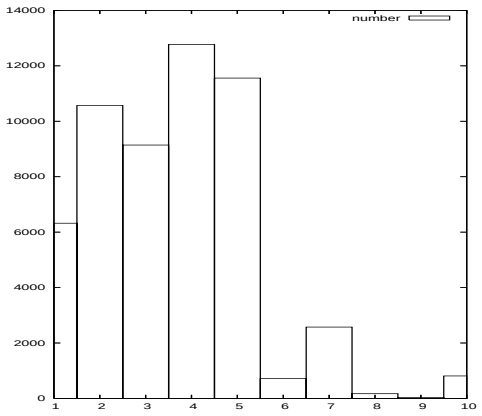



FIGURE 4.1 – Fréquences d'apparitions (nombre d'images en ordonnée) des différents personnages (identifiant en abscisse) sur les bandes sonore et visuelle sur une émission du jeu « Pyramide ».

### 4.1.2 Matrice de co-occurrences

La matrice de co-occurrences $m$ de taille $n_a * n_v$, où $n_a$ est le nombre total de locuteurs dans l'index audio et $n_v$ est le nombre total de personnes dans l'index vidéo, représente la valeur de l'intersection en termes de nombre d'images entre les index audio et vidéo :

$$m = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \dots & V_{n_v} \end{array} \\ \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1n_v} \\ m_{21} & m_{22} & \dots & m_{2n_v} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n_a1} & m_{n_a2} & \dots & m_{n_an_v} \end{pmatrix} \end{array} \tag{4.1}$$

avec :

- $\{A_i\}_{i=1\dots n_a}$ l'ensemble des locuteurs,
- $\{V_j\}_{j=1\dots n_v}$ l'ensemble des personnes visuelles,
- $m_{ij}$ correspond au nombre total d'images où la voix $A_i$ est entendue pendant que la personne $V_j$ apparaît.

La première intuition serait de trier les lignes (resp. colonnes) de cette matrice pour décider d'une association voix/image. Dans ce cas, nous supposerions que l'image associée à une voix $A_i$ est celle qui est la plus présente lorsque $A_i$ est entendue (resp. que la voix associée à une image $V_j$ est celle qui est la plus entendue lorsque $V_j$ apparaît). Or, ceci n'est pas toujours le cas !

Par exemple, dans un jeu TV, lorsqu'un candidat apparaît à l'image, la voix la plus entendue n'est pas obligatoirement la sienne, mais peut être celle du présentateur. La fréquence de parole du présentateur est très importante, en comparaison de celle des candidats alors que la visualisation de celui-ci reste limitée.

Néanmoins, il existe un « sens de lecture » qui produit une solution correcte. Par exemple, si la voix la plus fréquente lorsqu'un candidat apparaît n'est pas la sienne, dans ce cas la solution inverse est vraisemblablement correcte, c'est-à-dire que lorsque nous entendons sa voix, l'image la plus fréquente à l'écran doit être la sienne.

Pour déterminer un « sens de lecture », nous allons calculer deux nouvelles matrices, $m_a$ et $m_v$ de fréquences d'apparition des locuteurs $A_i$, resp. des personnes visuelles $V_j$ :

$$m_a = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \dots & V_{n_v} \end{array} \\ \begin{pmatrix} f^a_{11} & f^a_{12} & \dots & f^a_{1n_v} \\ f^a_{21} & f^a_{22} & \dots & f^a_{2n_v} \\ \dots & \dots & \dots & \dots \\ f^a_{n_a1} & f^a_{n_a2} & \dots & f^a_{n_an_v} \end{pmatrix} \end{array} \tag{4.2}$$

$$m_v = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \dots & V_{n_v} \end{array} \\ \begin{pmatrix} f^v_{11} & f^v_{12} & \dots & f^v_{1n_v} \\ f^v_{21} & f^v_{22} & \dots & f^v_{2n_v} \\ \dots & \dots & \dots & \dots \\ f^v_{n_a1} & f^v_{n_a2} & \dots & f^v_{n_an_v} \end{pmatrix} \end{array} \tag{4.3}$$

avec $f^a_{ij} = \frac{m_{ij}}{\sum\limits_j m_{ij}}$ et $f^v_{ij} = \frac{m_{ij}}{\sum\limits_i m_{ij}}$.

À partir de ces deux matrices, nous définissons la matrice de fusion $F$, en calculant pour chaque couple $(i, j)$ une valeur à partir des fréquences $f^a_{ij}$ et $f^v_{ij}$. Plusieurs opérateurs de fusion ont été examinés durant cette étude : maximum, moyenne, produit, etc.

$$F = \begin{pmatrix} C(A_1, V_1) & \dots & C(A_1, V_{n_v}) \\ C(A_2, V_1) & \dots & C(A_2, V_{n_v}) \\ \dots & \dots & \dots \\ C(A_{n_a}, V_1) & \dots & C(A_{n_a}, V_{n_v}) \end{pmatrix} \tag{4.4}$$

avec $C(A_i, V_j)$ le coefficient de fusion entre la voix $A_i$ et la personne $V_j$.

Suivant le besoin de l'application, il est possible d'associer plusieurs voix à une personne par une lecture en colonne ou d'associer plusieurs personnes à une voix par une lecture en ligne. Dans tous les cas, la valeur maximale correspond à la meilleure association.

Cette méthode a été évaluée sur le jeu télévisé « Pyramide ».

Dans la mesure où nous disposons d'une vérité terrain en termes de segmentation audio et vidéo, nous avons pu évoluer notre seule association automatique, fondée sur les matrices de co-occurrences : les résultats sont excellents. En effet, l'association entre le costume et la voix de chaque personnage est parfaite.

Dans le cas automatique, bien évidemment des problèmes de sur-segmentation des index audio et vidéo impactent les performances finales. Cependant, les résultats restent cohérents et les meilleurs scores sont obtenus par l'opérateur de fusion « produit ». Voici dans la table 4.1 un exemple de résultats où la segmentation vidéo extrait 19 costumes et la segmentation audio révèle 20 locuteurs. La vérité terrain indique 13 personnes en visuel dont 6 principales et 8 locuteurs dont 6 principaux, identiques au visuel. L'association automatique des segmentations en costumes et en locuteurs permet ici de détecter 6 personnages : P1 à P6. Dans cet exemple, les locuteurs 16 et 18 sont associés au costume 1 pour former le personnage P1.

TABLE 4.1 – Association audio-vidéo d'index automatiques par matrices de co-occurrences. 6 personnages sont détectés dans cette exemple par l'association du costume et de la voix.

| personnage | P1 | P2 | P3 | P4 | P5 | P6 |
|---|---|---|---|---|---|---|
| costume (vidéo) | 1 | 2 | 5,15 | 6 | 9,10 | 12 |
| voix (audio) | 16,18 | **2** | 4,6 | 1,3,5 | 7,8,10 | 12 |

Les erreurs sont de deux types. D'une part, les valeurs soulignées correspondent à des inversions de voix au sein d'une même équipe de deux joueurs. D'autre part en gras, la voix 2 est associée au costume 2 (une occurrence du présentateur) alors qu'il s'agit d'une voix de personne non visible.

Autant le premier type d'erreur doit se régler par une amélioration des méthodes de segmentation automatique (locuteurs et/ou costumes), autant le second mérite une attention particulière. C'est pourquoi nous nous sommes intéressés à la caractérisation de l'intervenant à travers sa présence ou son absence à l'écran : nous parlons de « voix IN », « voix OUT » et « voix OFF ».

## 4.2 Caractérisation des intervenants

Notre objectif est de savoir si, à un instant donné, sans connaissance *a priori* sur le type de document traité, un locuteur est visible ou non. Jusque là, les travaux pouvaient se résumer ainsi : un personnage est de classe *IN* lorsqu'il est détecté à l'écran pendant son élocution, sinon il est *OUT*. Toutefois, cette classification un peu arbitraire ne prend pas en considération l'activité visible de parole à part entière : la personne détectée à l'écran n'est pas forcément celle qui parle.

Cette première étude se place dans un contexte simplifié, à savoir que l'enregistrement vidéo se fait sur un même lieu.

Nous avons donc défini trois classes d'intervenants :
- la personne qui parle est visible, elle est de classe **IN**,
- la personne qui parle n'est pas visible, mais elle a déjà été filmée ou le sera durant son élocution, elle est de classe **OUT** ; elle est sur la scène d'enregistrement vidéo,
- la personne qui parle n'est jamais visible durant sa locution, elle est de classe **OFF** ; elle n'est pas sur la scène d'enregistrement.

### 4.2.1 Hypothèses de travail

Nous définissons un **segment audiovisuel** comme une séquence pendant laquelle une classe d'intervenant reste stable. Nous faisons l'hypothèse qu'un tel segment est délimité par une frontière correspondant soit à un changement de locuteurs, soit un changement de plans, soit une combinaison des deux ou un silence. Nous nous appuyons sur les travaux de l'équipe : la segmentation PMB (voir section I) pour la détection de silence, la SRL (voir section 1.1) pour isoler les locuteurs, et la détection des changements de plans [Jaffré 04].

Compte tenu de ce positionnement, trouver les segments audiovisuels revient à étudier l'ensemble des segments de base (changements de locuteurs, de plans, silence) et rechercher sur leurs frontières correspondent à un changement de classe d'intervenant.

Compte tenu de l'hypothèse faite sur l'unité de la scène visuelle, les changements $OFF \leftrightarrow IN$ et $OFF \leftrightarrow OUT$ correspondent à des changements d'environnement sonore. Nous nous sommes, dans un premier temps, attachés à détecter ces types de changements pour ensuite préciser les trois classes.

### 4.2.2 Changement d'environnement sonore

L'étude de la représentation cepstrale a montré tout l'intérêt du calcul de la moyenne à long terme des vecteurs cepstraux pour délimiter le signal de parole. Nous détournons cette propriété pour détecter un changement d'environnement : si un changement d'environnement sonore intervient, la soustraction cepstrale ne doit pas être réalisée sur l'ensemble de l'enregistrement mais sur les deux parties séparément pour une meilleure représentation.

Le test sur le changement d'environnement sonore est un test d'hypothèse fondé sur le maximum de vraisemblance : « Generalized Likelihood Ratio » (GLR, [Gish 91]) :

$$\begin{cases} h_0 = \text{L'environnement sonore est stable de part et d'autre de la frontière considérée} \\ h_1 = \text{L'environnement sonore est instable} \end{cases} \tag{4.5}$$

Un segment de 4 secondes est centrée sur chaque frontière potentielle et analysé pour fournir 250 vecteurs cepstraux de part et d'autre, soit $(y_i, i = 1, \ldots, 250)$ et $(y_i, i = 251, \ldots, 500)$.

Ce qui se traduit par :

$$\begin{cases} h_0 = \text{L'environnement sonore est caractérisé par un seul vecteur moyen cepstral } m_0 \\ h_1 = \text{L'environnement sonore est caractérisé par deux vecteurs calculés de part et d'autre} \\ \text{de la frontière } m_1 \text{ et } m_2 \end{cases} \tag{4.6}$$

Ou encore :

$$\begin{cases} h_0 = (y_1, \ldots, y_{500}) \text{ suit une loi } \mathcal{N}(m_0, \sum_0) \\ h_1 = (y_1, \ldots, y_{250}) \text{ suit une loi } \mathcal{N}(m_1, \sum_1) \text{ et } (y_{251}, \ldots, y_{500}) \text{ suit une loi } \mathcal{N}(m_2, \sum_2) \end{cases} \tag{4.7}$$

Le rapport de vraisemblance s'écrit :

$$\Delta_{0,1} = \frac{P(y_1, \ldots, y_{500}/h_0)}{P(y_1, \ldots, y_{500}/h_1)} = \frac{P(y_1, \ldots, y_{500}/\mathcal{N}(m_0, \sum_0))}{P(y_1, \ldots, y_{250}/\mathcal{N}(m_1, \sum_1)) * P(y_{251}, \ldots, y_{500}/\mathcal{N}(m_2, \sum_2))} \tag{4.8}$$

En fixant un seuil, il est possible alors de prendre une décision en faveur d'une des deux hypothèses. Nous avons introduit la variable $\Delta_{t,t+1}$, booléen indiquant la stabilité ou non de l'environnement sonore du segment $t$ au segment $t+1$.

### 4.2.3 Système de classification IN-OUT-OFF

Il en résulte un système de classification original tant du point de vue de la paramétrisation que du module de décision. Nous rappelons que nous cherchons à regrouper des segments de base (changements de locuteurs, de plans, silence) en un segment audiovisuel (IN, OUT ou OFF).

**Paramétrisation**
Afin de caractériser au mieux chacun des segments de base, nous effectuons des traitements sonores et visuels.

**Côté vidéo**, nous réalisons une détection de visages par OpenCV, un indicateur booléen $P_t$ traduit cette présence ou absence de visage.

Sur deux images successives, nous calculons la différence (erreur quadratique moyenne) des pixels dans la zone des lèvres sur le canal de la luminance (espace HLS) et nous obtenons le *Taux d'Activité Labiale* (TAL).

En général, lorsqu'une personne parle, elle bouge plus que les lèvres. Sur le même principe, nous avons ainsi proposé deux autres mesures :
- le *Taux d'Activité du Visage* (TAV),
- le *Taux d'Activité du Corps* (TAC).

Il existe une hiérarchie entre les descripteurs plaçant dans l'ordre décroissant d'importance le TAL, le TAV, puis le TAC. Ainsi, les trois taux d'activité sont combinés en une seule fonction $\Phi(i,j)$ qui représente l'activité relative entre les segments $i$ et $j$. Un score positif indique qu'il y a plus d'activité dans le segment $i$ que $j$, il est incrémenté de la manière suivante :
- si $TAL(i) > TAL(j)$ alors $\Phi(i,j) = +3$ sinon $\Phi(i,j) = -3$,
- si $TAV(i) > TAV(j)$ alors $\Phi(i,j) = +2$ sinon $\Phi(i,j) = -2$,
- si $TAC(i) > TAC(j)$ alors $\Phi(i,j) = +1$ sinon $\Phi(i,j) = -1$.

**Côté Audio**, nous utilisons la variable booléenne $\Delta_{t,t+1}$, définie au paragraphe précédent.

Au final, si $t$ représente l'indice de segment de base, le récapitulatif des descripteurs sonores et visuels utilisés est le suivant :

- $P_t$ : booléen indiquant la présence (ou l'absence) de visage durant le segment $t$,
- $\Phi_{t,t+1}$ : score d'activité visuelle entre deux segments consécutifs $t$ et $t+1$,
- $\Delta_{t,t+1}$ : booléen indiquant la stabilité ou non de l'environnement sonore du segment $t$ au segment $t+1$,
- $T_{t,t+1}$ : type de transition audio et/ou vidéo $\in \{S, L, SL\}$ d'un segment $t$ à $t+1$ ($S$ pour un changement de plan, $L$ pour changement de locuteur et $SL$ pour les deux).

**Module de décision**

Nous avons choisi comme classifieur un automate à états finis : **IN**, **OUT**, **OFF**, ainsi qu'un état de **REJET**. Celui-ci nous sert d'état initial ainsi que d'échappatoire lorsque les informations dont nous disposons ne sont pas suffisantes pour attribuer une classe à un intervenant (voir figure 4.2).

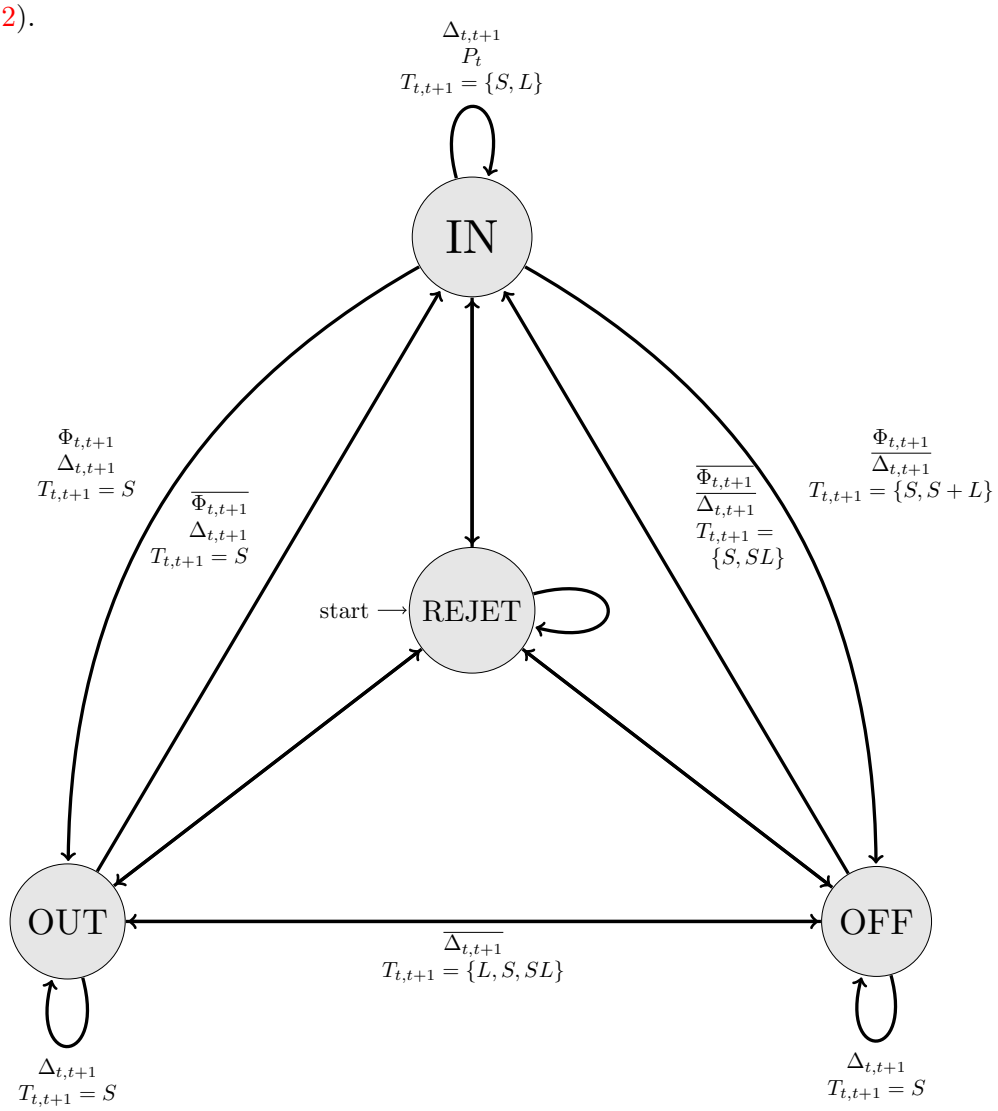



FIGURE 4.2 – Automate de caractérisation des segments audiovisuels des intervenants.

Ce travail de classification IN/OUT/OFF a été évalué sur un enregistrement de journal télévisé du corpus TRECVID2004 [Kraaij 04] et sur l'émission de Pyramide utilisée pour la détection des intervenants (cf. section précédente).

La classe **REJET** est choisie dans 24,2 % des cas. Si nous considérons cette classe comme une classification correcte, nous obtenons une accuracy de 87,1 % sinon le score chute à 55,8 %, Il est intéressant de noter que si nous ne considérons que les segments qui ne sont pas classés **REJET**, nous obtenons une accuracy de 82,6 %.

Après avoir détecté les intervenants et les avoir caractérisés en fonction de leur présence, nous allons étudier leurs interactions dans un objectif de structuration à l'échelle du document.

## 4.3 Structuration en programmes

Nous souhaitons utiliser l'intervenant, et plus précisément le locuteur, comme élément de structure des documents audiovisuels.

En règle générale dans les documents diffusés, les **séquences d'interaction entre intervenants** correspondent à de nombreux événements tels que des interviews, des débats, des tours de jeu, etc. Ces séquences font généralement parties d'une ligne directrice de la part de la production. Par exemple, lors d'une émission politique, les producteurs peuvent décider de la nature des interventions (interactive ou non) et ainsi typer le programme (débat ou interview). D'ailleurs, la présence d'une interview dans un programme peut également mettre en évidence des séquences traitant d'actualités importantes. Ces séquences d'interaction ne sont généralement pas indiqués dans les guides de programmes, d'où un certain intérêt de les détecter automatiquement pour cibler les zones d'informatives.

De la même façon, le **rôle des intervenants** peut être fondamental. Les gens apparaissent généralement dans un programme avec un rôle bien défini qui peut influer directement sur la façon dont ils parlent au public ou aux autres intervenants sur le plateau. Par exemple, la plupart des émissions diffusées sont menées par un présentateur dont la responsabilité peut être très importante : il est en charge de l'ouverture et de la clôture de l'émission, il présente les sujets et les invités, il gère les prises de parole, etc.

Partant de ces constats et tirant profit d'une détection automatique quasi-parfaite du présentateur (voir section 1.3), nous privilégions le présentateur pour rechercher des éléments structurants.

### 4.3.1 Structurations primaire et secondaire

Afin de décrire les unités logiques de structuration, 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 définit une séquence correspondant à une sous-partie de programme, éventuellement à un programme complet, que nous appelons* ***brique élémentaire*** ».

Ce postulat nous conduit à définir deux types d'unités de structuration, illustrées sur la figure 4.3 :

- les **unités présentées** sont les briques élémentaires bornées maximales d'un document,
- 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...

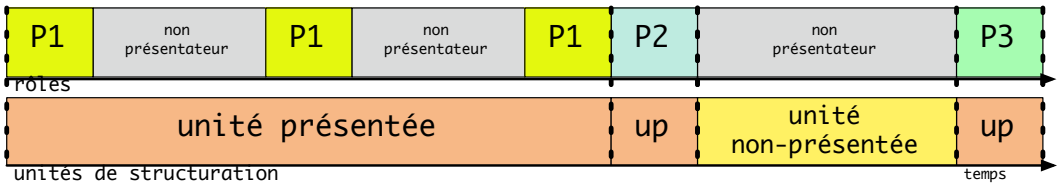



FIGURE 4.3 – Exemple d'unités de structuration.

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 structurations.

Rappelons que les rôles sont *Présentateur*, *Journaliste* et *Autre* (cf. section 1.3). Soit $d(journaliste)$ le temps de parole cumulé de tous les journalistes de l'unité, et $d(autre)$ le temps de parole cumulé des intervenants qui ne sont ni présentateur, ni journaliste.

La **structuration primaire** assigne les unités présentées dans trois catégories, en rapport avec le temps de parole imputable à chaque type de rôle présent sur cette unité (cf. figure 4.4) :

- les **informations** sont les unités de programmes sur lesquelles : $d(journaliste) \leq d(autre)$,
- les **entretiens** correspondent aux zones où : $d(autre) \leq d(journaliste)$,
- la **transition** contient uniquement un présentateur.

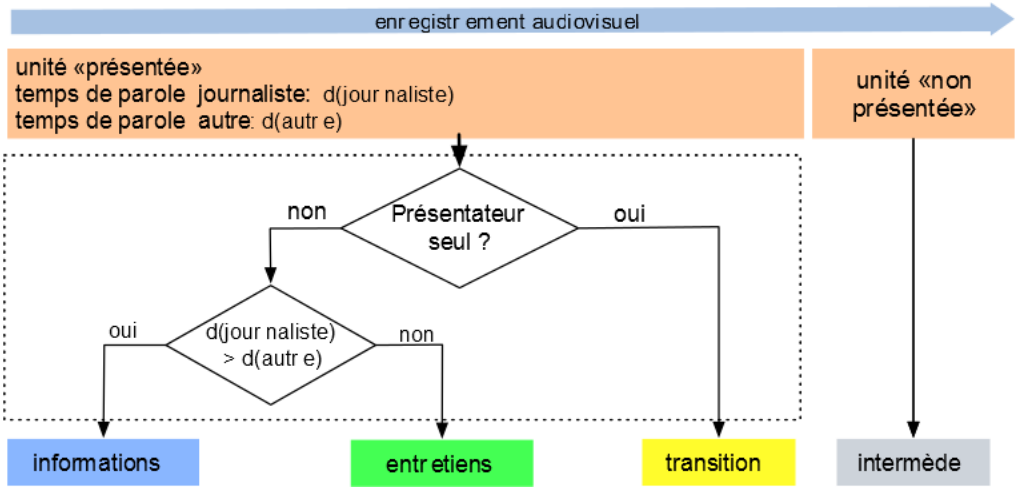



FIGURE 4.4 – Structuration primaire des émissions.

Un second postulat nous introduit la **structuration secondaire** : « *dans un programme audiovisuel, les conversations entre les intervenants sont liées à des choix de la production. Les rôles impliqués dans les interactions apportent une information sur la nature de la séquence conversationnelle* ».

Ainsi, la prise en compte des zones d'interaction (z.i., voir section 1.2) dans les unités d'informations et d'entretiens, peut en fonction du contexte correspondre à des types différents d'interactions (voir figure 4.5) :

- l'**interview** est une z.i. entre un présentateur et un interviewé,
- la **chronique** est une z.i. impliquant un présentateur et un journaliste,
- le **débat** est une z.i. entre deux invités, ou entre un invité et un journaliste,
- le **relais** est une z.i. entre deux présentateurs.

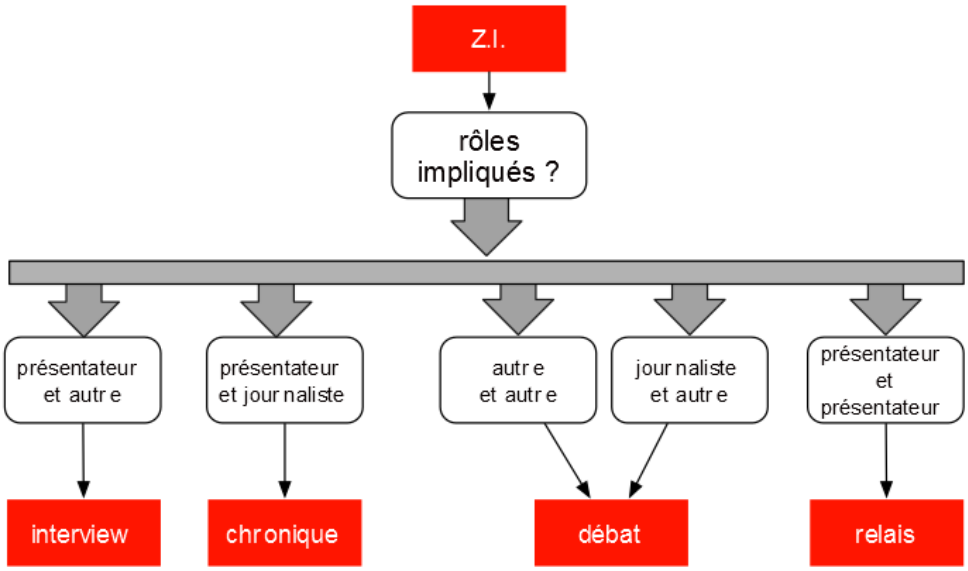



FIGURE 4.5 – Structuration secondaire des émissions.

### 4.3.2 Validation des structurations primaire et secondaire

Notre système complet de structuration en émissions est illustré dans la figure 4.6. Rappelons qu'il est fondé sur les locuteurs, en l'occurrence la SRL présentée en section 1.1. À la suite de celle-ci, deux traitements sont effectués en parallèle : la caractérisation des z.i. vue dans la section 1.2, la reconnaissance automatique des rôles décrite dans la section 1.3 et l'ensemble des règles décrites dans la section précédente 4.3.1.

Sur le corpus de test du projet EPAC [10], composé de 10 heures, le DER de la SRL est de 7 %, nous obtenons des z.i. d'un niveau compris entre 1 et 14 et le système de reconnaissance à 5 rôles obtient un score de reconnaissance de 92 %.

10. http://projet-epac.univ-lemans.fr/

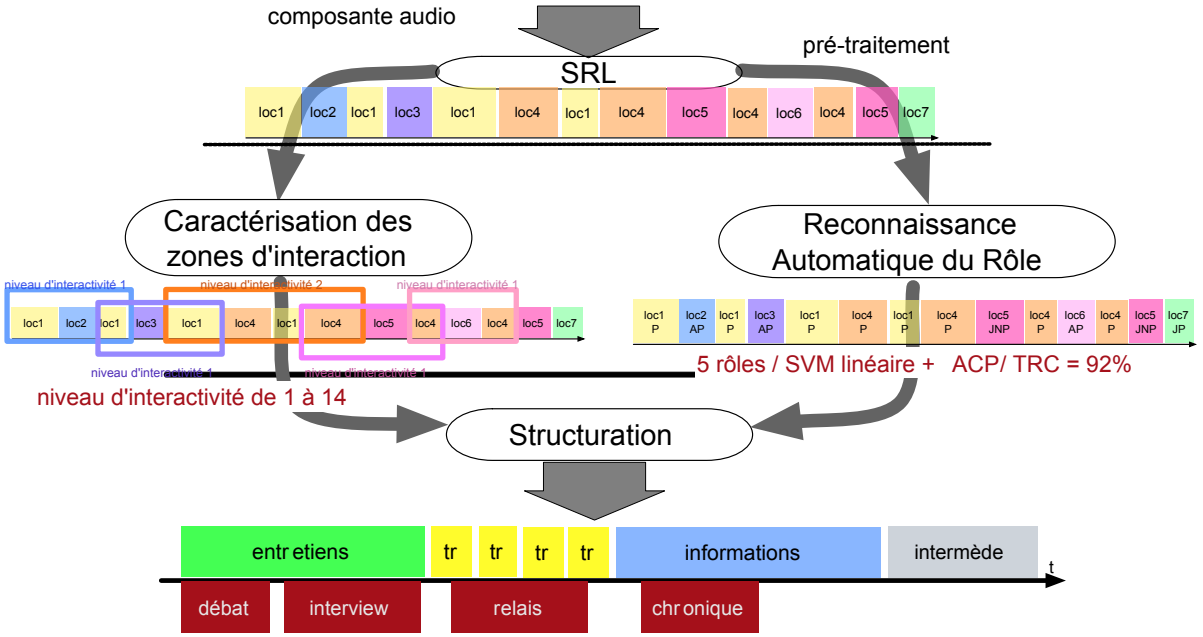



FIGURE 4.6 – Système de structuration des émissions fondé sur les intervenants.

En utilisant ces différents systèmes, nous obtenons un taux de reconnaissance (la durée correctement classée sur la durée totale) de 85,6 % pour la structuration primaire. Quelques confusions existent entre les unités « entretiens » et « informations » à cause des erreurs du système de reconnaissance des rôles (inversion entre les rôles *autre* et *journaliste*).

Les résultats de la structuration secondaire sont de 67,1%. Ce score est beaucoup moins bon pour principalement deux raisons. D'une part, les rôles *journaliste* et *autre*, qu'ils soient ponctuels ou non, influent beaucoup plus dans cette étape de structuration. D'autre part, les zones d'interaction servant de base à ce niveau, sont extrêmement sensibles aux erreurs de segmentation locales introduites par la SRL, en l'occurrence les zones de parole superposées à de la musique.

Ce travail, sans connaissance *a priori* et grâce au pouvoir d'ancrage du présentateur, effectue une macro-segmentation cohérente (premier niveau) à la manière d'une grille de programme, ainsi qu'une micro-segmentation (second niveau) qui met en évidence des zones caractéristiques du contenu : les séquences conversationnelles. Cette approche est assez spécifique des contenus formatés, tels ceux de la télévision et la radio, où la production établit des règles de structure claires afin de cadrer et guider l'auditeur (ou le téléspectateur). Le contenu de vidéo-surveillance ou de vidéos personnelles ne rentrant pas dans ce schéma là, ceci nous a poussés à proposer une approche plus générale pour structurer des enregistrements personnels. Cependant, nous nous sommes restreints à la structuration de la journée d'une personne en fonction des activités qu'elle effectue.

## 4.4 Segmentation en activités

Ce travail de segmentation ou structuration en activités de la vie quotidienne repose sur une collaboration forte entre trois entités : le LaBRI [11], l'IMS [12] et l'IRIT [13], et plus précisément trois personnes qui sont Jenny Benois-Pineau, Rémi Mégret et moi-même.

À travers différents projets (PEPS CNRS, ANR Blanc), nous avons développé des outils et des méthodes pour l'indexation des Activités de la Vie Quotidienne (AVQ) dans les vidéos acquises par une caméra portée par un patient à son domicile. Le contexte applicatif de ce travail est l'étude des difficultés fonctionnelles d'une personne dans l'accomplissement des activités instrumentales de la vie quotidienne car celles-ci peuvent apparaître jusqu'à dix ans avant l'établissement d'un diagnostic clinique des maladies de démence, basé sur les méthodologies cognitives de référence [Pérès 08].

L'un des principaux verrous au niveau du traitement automatique de telles données est de gérer la masse d'information capturée de façon à fournir au médecin une visualisation efficace pour l'analyse médicale des activités. Il est indispensable d'organiser le flux audio-vidéo en détectant les zones d'intérêt pour le médecin.

Cette problématique s'inscrit parfaitement dans la thématique de l'indexation multimédia. Cependant, en comparaison des contenus habituellement traités, les vidéos acquises par la caméra embarquée possèdent un contenu beaucoup plus variable (mouvement, luminosité, saturation, flou, bruits, etc.) : ceci complique d'autant plus leur analyse automatique.

Notre défi a été de proposer des paramètres, des modèles et des algorithmes adaptés à ce type de données.

### 4.4.1 Descripteurs audio-vidéo

Nous avons combiné des descripteurs vidéo et audio, organisés en 3 modalités : mouvement, visuel et sonore développées respectivement par le LaBRI, l'IMS et l'IRIT.

Des paramètres originaux ont émergé de ce travail :

- concernant le **mouvement**, trois nouveaux descripteurs ont ainsi été développés afin de caractériser l'historique du mouvement, sa force et le mouvement résiduel. Ils capturent à la fois le mouvement global et la présence d'objets tels que les mains,
- au niveau **visuel**, la localisation de la pièce où se trouve la personne a été estimée grâce à un descripteur « moyen-niveau » obtenu par une classification du contenu visuel utilisant la fusion de caractéristiques visuelles multiples.
- pour la modalité **sonore**, deux descripteurs « bas-niveau » ont été spécialement créés dans cette étude.

Pour les deux premières modalités, ne s'agissant pas de mon travail, je renvoie aux thèses suivantes : [Dovgalecs 11] et [Karaman 11].

11. http://www.labri.fr/
12. http://www.ims-bordeaux.fr/
13. http://www.irit.fr/

La particularité de ma contribution consiste en l'utilisation de descripteurs sonores de « bas-niveau ». Dans l'environnement de la maison, il y a beaucoup de sons significatifs : téléviseur, radio, le bruit des objets manipulés, les appareils d'électroménager, les conversations avec les personnes, etc. Tous ces sons sont de bons indicateurs de l'activité effectuée par la personne et de sa localisation. Afin de décrire une partie de l'environnement sonore, différents jeux de paramètres sont extraits. Chaque jeu est représentatif d'un son particulier (parole, musique, bruit et silence) avec un objectif constant « avoir des caractéristiques sonores robustes » car d'une maison à l'autre les conditions sonores sont très hétérogènes. L'ensemble du système audio est illustré sur la figure 4.7 : celui-ci ne prend pas de décision mais retourne un score de confiance pour chacun des sept détecteurs.

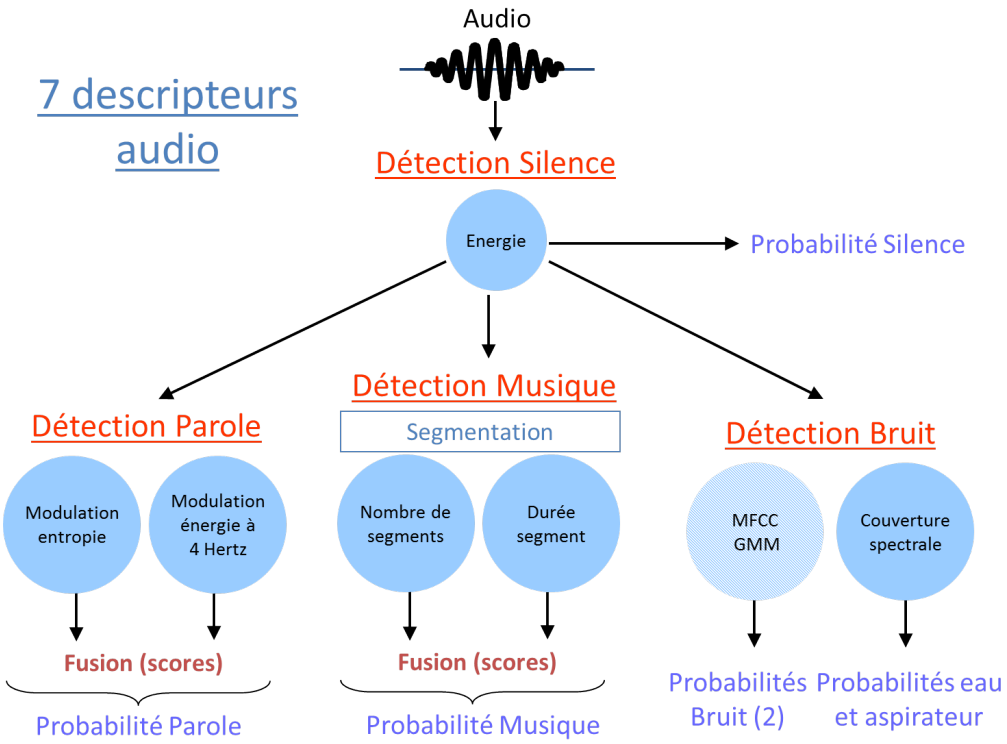



FIGURE 4.7 – Schéma d'extraction des 7 descripteurs sonores.

Le premier paramètre est un classique détecteur d'activité vocale fondé sur l'énergie du signal. Ensuite, une utilisation du système de segmentation PMB de la section I permet d'extraire deux autres descripteurs indiquant la présence plus ou moins sûre de parole et de musique.

Compte tenu du contexte applicatif, la classe bruit est représentée de manière plus fine à travers 4 paramètres. Les deux premiers descripteurs sont issus du système présenté dans la section 3.2.1 pour les applaudissements et les rires afin d'avoir une idée de la nature percussive et périodique des sons. Les deux autres descripteurs sont fondés sur la couverture spectrale (voir section 3.2.2). Suivant le seuil utilisé, nous avons une indication sur la présence d'eau et d'aspirateur.

Les 60 descripteurs vidéos (dynamiques et statiques) et audio sont répertoriés dans le tableau 4.2) ; ils vont alimenter notre système de segmentation en activités.

TABLE 4.2 – Liste des descripteurs audio-vidéo pour la reconnaissance d'activités.

| Modalité | Notation | Description | Nombre |
|---|---|---|---|
| Dynamique | $H_{tpe}$ | Mouvement global instantané | 10 |
| | $H_c$ | Historique du mouvement global | 8 |
| | $RM$ | Résidu du mouvement local | 16 |
| Statique | $CLD$ | Color Layout Descriptor (MPEG-7) | 12 |
| | $Loc$ | Localisation | 7 |
| Sonore | $Audio$ | PMB | 7 |

### 4.4.2 Système de segmentation en activités

Le système de base a été développé au LaBRI [Karaman 10] : il implique un Modèle de Markov Caché Hiérarchique (HHMM) à deux niveaux.

Les activités de la vie quotidienne sont codés dans le niveau 1 du HHMM. L'ensemble des états possibles est défini selon la nomenclature des types de comportement, côté médical.

L'ensemble des activités considérées contient la plupart des activités de l'analyse IADL (Instrumental Activities of Daily Living) que les praticiens utilisent actuellement dans les enquêtes papier. Voici la liste des 24 activités : « Passer l'aspirateur », « Passer le balai », « Faire la lessive », « Servir un plat/verre », « Préparer le café », « Faire la cuisine », « Se brosser les cheveux », « Téléphoner », « Regarder la télévision », « Tricoter », « Jardiner », « Écouter la radio », « Essuyer la vaisselle » « Se brosser les dents » « Faire la vaisselle », « Faire le lit », « Lire », « Utiliser un ordinateur », « Prendre des médicaments », « Payer », « Utiliser une machine », « Discuter », « Se laver les mains » et « Se déplacer ».

Un état de rejet est également présent dans le HHMM pour modéliser les observations non significatives du point de vue des médecins. Ainsi défini, le niveau 1 contient les transitions entre des activités de type « sémantique ». La matrice de transition est fixée *a priori*, en fonction de l'environnement familial du patient et toutes les probabilités initiales sont égales.

Le niveau 2 du HHMM représente chaque activité par 3 états. Les états sont modélisés par un mélange de 5 lois gaussiennes. L'apprentissage des lois et de la matrice de transitions s'effectuent par l'algorithme de Baum Welsh [Rabiner 89].

J'ai pris part à l'étude de différentes méthodes de fusion (précoce, intermédiaire et tardive) afin d'exploiter au mieux l'ensemble des caractéristiques multimodales disponibles (60 descripteurs) et plusieurs systèmes de reconnaissance d'activités en sont dérivés.

La **fusion précoce** consiste à concaténer les descripteurs préalablement à la classification. La dimension de l'espace de description est de 60 lorsque tous les descripteurs sont utilisés. Différentes configurations (combinaisons de descripteurs) ont été testées [Karaman 10].

La **fusion intermédiaire** traite les différentes modalités séparément. Contrairement à la fusion précoce, l'observation n'est pas la concaténation de tous les descripteurs, mais un ensemble de trois flux d'observations correspondant au nombre de modalités.

Plus formellement, une observation $o_i \in \mathbb{R}^N$ est composé de $K$ modalités. Nous avons donc $K$ flux d'observations $o_{i,1}, ..., o_{i,k}$ avec $o_{i,k} \in \mathbb{R}^{N_k}$ et $\sum_k N_k = N$.

Pour chaque état du HHMM les observations de chaque flux sont modélisées séparément par un mélange de lois gaussiennes. Chaque flux $k$ est pondérée par un poids $w_{lk}$ qui dépend de l'activité $l$, sachant que le poids est le même pour tout état de la même activité avec la contrainte $\sum_k w_{lk} = 1$. La probabilité d'observation $o_i$ pour l'état $q_j$ est :

$$p(o_i/q_j) = \prod_{k=1}^{k} p_k(o_{i,k}/q_j)^{w_{lk}} \tag{4.9}$$

avec $p_k(o_{i,k}/q_j)$ la probabilité de l'observation partielle $o_{i,k}$ du flux $k$ pour le kième mélange de lois gaussiennes associé à l'état $j$.

Remarque : chaque état est composé de 3 flux (audio, statique et dynamique) dont les poids sont équiprobables d'après notre étude [Karaman 12], $w_{lk} = \frac{1}{3}$.

La **fusion tardive** utilise un HHMM pour chaque modalité (dynamique, statique et audio) et fusionne les résultats de ces classificateurs préliminaires pour la prise de décision. Chaque modalité $k$ a un score de confiance qui est spécifique à chaque activité $l$. Celui-ci est calculé à partir du taux d'accuracy, ou performance $\text{perf}_{lk}$, obtenu par la modalité $k$ sur l'activité $l$ :

$$e_{lk} = \frac{\text{perf}_{lk}}{\sum_k \text{perf}_{lk}} \tag{4.10}$$

La décision finale est calculée ainsi : pour chaque observation $o_i$, nous choisissons l'activité $l_i$ qui a le meilleur score de confiance parmi toutes les modalités :

$$l_i = arg\,max(e_{liAudio}, e_{liDynamique}, e_{liStatique}) \tag{4.11}$$

Les trois méthodes de fusion ont été testées sur une partie du corpus du projet IMMED [14], à savoir 14 heures provenant de 34 personnes différentes, donc 34 lieux différents.

L'évaluation s'effectue par une validation croisée « leave-one-out ». Les résultats de chaque approche de fusion sont présentés dans le tableau 4.3, en terme d'accuracy par rapport à la durée du corpus de test.

14. http://immed.labri.fr/

TABLE 4.3 – Performances des fusions précoce, intermédiaire et tardive pour la segmentation en 24 activités du quotidien.

| Métrique d'évaluation | Fusion | | |
|---|---|---|---|
| | précoce | intermédiaire | tardive |
| Accuracy | 20,7 % | 44,2 % | 21,5 % |

La fusion intermédiaire surclasse les autres, avec une accuracy de 44,2 %. Il est intéressant de noter que chaque fusion est meilleure que les modalités prises individuellement (la meilleure étant l'audio avec une accuracy de 11,1 %).

Nous venons d'aborder dans cette partie des segmentations audiovisuelles fondées directement sur l'intervenant. Ces segmentations ont montré tout leur potentiel en obtenant des résultats assez convaincants. Ceci est assez logique car les méthodes de base (segmentations en locuteurs et en visages) ont maintenant une bonne maturité. En effet, comme elles représentent les maillons essentiels de toute indexation sonore et visuelle, elles ont été grandement étudiées par la communauté scientifique.

Nous allons maintenant traiter de l'organisation du contenu audiovisuel dans son ensemble : l'objectif est la structuration de documents.

# Chapitre 5

# Segmentation autour de la similarité

## Sommaire


**5.1 Similarité des documents . . . . . 76**
5.1.1 L'Intersection Quadratique Récursive . . . . . 76
5.1.2 Application des matrices de similarité aux contenus sonores . . . . . 78
**5.2 Chapitrage . . . . . 80**
5.2.1 Méthode de chapitrage . . . . . 80
5.2.2 Résultats de « chapitrage » . . . . . 83
**5.3 Organisation de contenus audiovisuels . . . . . 85**
5.3.1 Système d'organisation de contenus audiovisuels . . . . . 86
5.3.2 Interface du système d'organisation de contenus audiovisuels . . . . . 90


Dans le contexte de la structuration, mon travail se focalise sur l'organisation aussi bien intra-document qu'inter-documents. Cette démarche est réalisée en fonction du contenu de chaque document. Le choix de descripteurs adéquats est déterminant. D'une part, ils doivent permettre de représenter le contenu des zones homogènes et/ou la nature des transitions du document. D'autre part, ils doivent représenter les informations qui permettent la comparaison entre documents, via une distance ou une similarité.

Plutôt que d'utiliser le terme « structuration » dans le titre qui semblerait au premier abord mieux correspondre à des notions d'agencement de document et de mise en forme, j'ai préféré le mot « segmentation ».

La première raison est liée à l'essence même de ma position développée dans ce manuscrit, à savoir que tous les résultats sont issus de segmentations *a priori* ou *a posteriori*.

La deuxième motif est qu'en utilisant le mot « structuration », cela sous-entend que le document possède une structure. Or celle-ci n'est pas forcément présente, ou en tous cas pas directement accessible. Les vidéos personnelles en sont un criant exemple car la notion de structure est plus conforme à des enregistrements maîtrisés : télévisuels, radiophoniques, studio, etc.

La dernière explication est qu'un document est multi-échelle et qu'utiliser le terme de « structuration » sous-entend qu'il est traité sur différents niveaux simultanément alors qu'en général il s'agit d'un traitement spécifique pour chaque niveau. D'ailleurs, nous utilisons les termes de « segmentation en plans » pour la vidéo et de « segmentation en locuteurs » pour l'audio...

Être capable de dire que deux documents audiovisuels sont identiques, se ressemblent un peu, beaucoup, ou sont complètement différents n'est ni utopique, ni subjectif, même si les contenus audio-vidéo sont néanmoins très hétérogènes et très complexes. En effet, comment comparer la chanson *Laura* et le film *Terminus* ? Peut-on dire qu'ils sont proches, parce qu'il y a le même interprète, ou parce qu'il s'agit de la même année de réalisation, ou autre chose...

Dans cette section, nous allons comparer des documents (ou des parties de document) à l'aide de distances entre descripteurs « bas-niveau ». Ainsi, la similarité intra et inter-documents est présentée ici à travers trois études.

La premier travail concerne une application des matrices de similarité au contexte sonore : il a été réalisé par Ali Mcheik lors de son stage de Master 2 [Mcheik 06].

La généralisation de cette approche a servi de base à un « chapitrage » universel de contenus audio-vidéo lors du projet ANR RIAM Chapitre par le post-doctorant Zein Al Abidin Ibrahim.

La dernière étude va au-delà de la notion de structure en proposant une organisation individuelle de documents, prenant en compte les choix (et les goûts) d'un utilisateur. Ce travail fut l'objet de la thèse CIFRE de Jérémy Philippeau [Philippeau 09] avec l'INA [15] sous le co-encadrement de Jean Carrive.

Ces trois travaux ont tous été effectués en co-encadrement avec Philippe Joly.

## 5.1 Similarité des documents

Pour évaluer la similarité entre deux documents, il faut rechercher les éléments communs. Chaque document est représenté par l'ensemble de ses caractéristiques. Ces caractéristiques peuvent être vues comme des séries chronologiques, le problème se ramène alors à la recherche de séquences similaires entre les deux séries.

Afin de détecter toutes les paires de séquences semblables, Siba Haidar a proposé une méthode, appelée *Intersection Quadratique Récursive* (IQR) lors de sa thèse encadrée par Philippe Joly. Ce travail a été utilisé pour comparer des documents vidéo, en se fondant uniquement sur des descripteurs visuels. Un descripteur est jugé pertinent s'il est corrélé à l'évolution temporelle du contenu et s'il peut être représenté par une valeur numérique sur la fenêtre d'analyse étudiée.

Ce travail étant le point de départ de notre étude, j'en rappelle brièvement le principe dans la section suivante.

### 5.1.1 L'Intersection Quadratique Récursive

Sachant qu'à l'échelle du document chaque descripteur génère une suite de valeurs, le principe est le suivant pour chaque descripteur :

– une première étape « grossière » consiste à ne conserver que les sous-séquences susceptibles d'être similaires,

15. http://www.ina.fr

– la seconde étape effectue une analyse « fine » et calcule le *taux de couverture* entre chaque sous-séquence,
– les scores ainsi obtenus sont stockés dans une matrice, dite de « similarité ».

Plus précisément, pour deux séquences $I$ et $J$ de même longueur à comparer, l'**algorithme IQR** se déroule de la manière suivante :

1. si les deux séquences sont similaires, i.e. $I \cap J \neq \emptyset$, et si la longueur des séquences est supérieure à $t_{max}$ alors les séquences sont coupées en deux sous-séquences de longueurs égales, soit $I$ en $I_1$ et $I_2$, et $J$ en $J_1$ et $J_2$ sinon fin de l'algorithme.
2. une comparaison quadratique est effectuée, c'est-à-dire une comparaison entre chacune des sous-séquences $I_1$ avec $J_1$ puis avec $J_2$, et $I_2$ avec $J_1$ puis avec $J_2$ (cf. figure 5.1). Il s'agit d'un retour à l'étape 1 pour chaque couple à comparer.

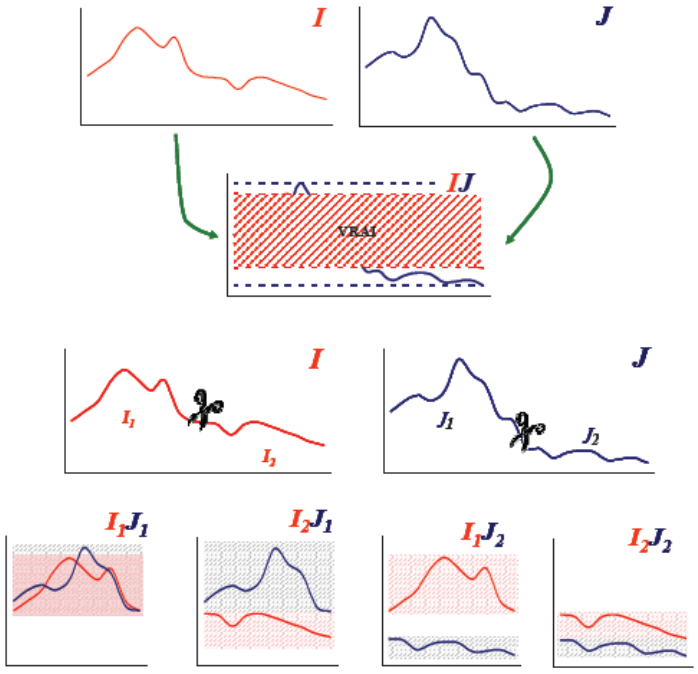



FIGURE 5.1 – Exemple de comparaison par l'algorithme IQR. Notons que l'algorithme continu seulement pour le couple $I_1J_1$ qui possède une intersection non vide.

Dans le cas où les valeurs des séquences comparées ont une intersection vide, la comparaison s'arrête : ceci évite des opérations de comparaison inutiles. Les couples de séquences identifiés par cet algorithme sont les séquences potentiellement similaires ou pouvant contenir des sous-séquences similaires. L'extraction des séquences similaires est fondé sur le taux de couverture.

Le **taux de couverture** de deux séquences $I$ et $J$ est le pourcentage du nombre de couples d'éléments ordonnés de $I$ et $J$ appariés. Un appariement optimal peut être obtenu par l'algorithme de programmation dynamique (DTW, Dynamic Time Warping). Sur l'exemple de la figure 5.2, le taux de couverture est de $10/16 * 100$.

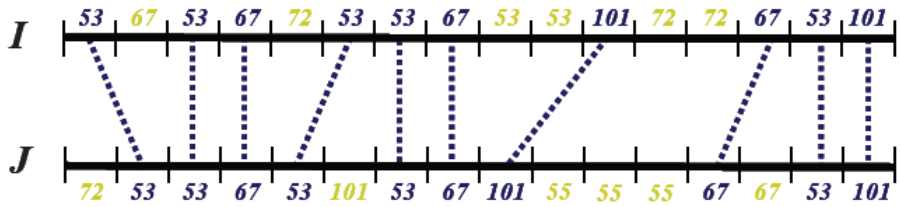



FIGURE 5.2 – Appariement entre deux séquences $I$ et $J$ de longueur 16. 10 appariements sont possibles ici.

La fonction du taux de couverture est calculée récursivement sur la taille $t_I$, $t_I = t_J$, des séquences $I$ :

$$\text{couv}(I,J) = \begin{cases} 0 & \text{si } I \cap J = \emptyset \\ 100 & \text{si } I \cap J \neq \emptyset \text{ et } t = 1 \\ \frac{1}{2}\max \begin{pmatrix} \text{couv}(I_1,J_1) + \text{couv}(I_2,J_2), \\ \text{couv}(I_1,J_2), \\ \text{couv}(I_2,J_1) \end{pmatrix} & I \cap J \neq \emptyset \text{ et } t > 1 \end{cases} \tag{5.1}$$

Ces scores de couverture sont ensuite stockée dans une **matrice de similarité** qui représente, pour un descripteur donné, le pourcentage de similarité entre les différents segments (séquences) de chaque document.

La fusion des descripteurs s'effectue en moyennant les matrices de similarité. Cette méthode a donné d'excellents résultats aussi bien pour de l'autosimilarité (similarité intra-document) que pour de la similarité entre documents à partir de 12 descripteurs visuels tels l'intensité lumineuse moyenne, les couleurs dominantes (espace HSV), le contraste, la quantité de mouvement, etc.

Pour plus de détails sur cette méthode, je vous conseille la lecture de la thèse de Siba Haidar [Haidar 05].

### 5.1.2 Application des matrices de similarité aux contenus sonores

En se fondant sur la méthode décrite précédemment et issues de descripteurs visuels, nous avons étudié spécifiquement sa généralisation à des descripteurs sonores.

Pour chaque trame d'analyse (en l'occurrence 16 ms) nous avons extrait 9 descripteurs « bas-niveau » :

- des paramètres temporels classiques que sont le taux de passage par zéro (ZCR) et l'énergie à court terme,
- des paramètres fréquentiels classiques, tels le centroïde spectral, le flux spectral et le spectral rollof point,
- mes quatre paramètres PMB, à savoir la modulation de l'énergie à quatre Hertz, la modulation de l'entropie, le nombre et la durée des segments (cf. section I).

Comme pour la vidéo, tous les paramètres n'ont pas la même pertinence suivant le contenu ciblé, nous verrons d'ailleurs dans l'étude suivante (section 5.2) comment sélectionner les plus pertinents. Pourtant les résultats sonores que nous obtenons sont très intéressants et en accord ou en complément des résultats visuels.

La figure 5.3 est un exemple de matrice de similarité calculée sur deux plages publicitaires à partir de la fusion de la modulation de l'énergie à quatre Hertz et de la modulation de l'entropie. Plus les coefficients sont élevés, plus les couleurs de la matrice sont chaudes. Sur la matrice de similarité de gauche, il est possible de voir des zones très différentes (rectangles bleus) et des zones similaires (traits rouges).

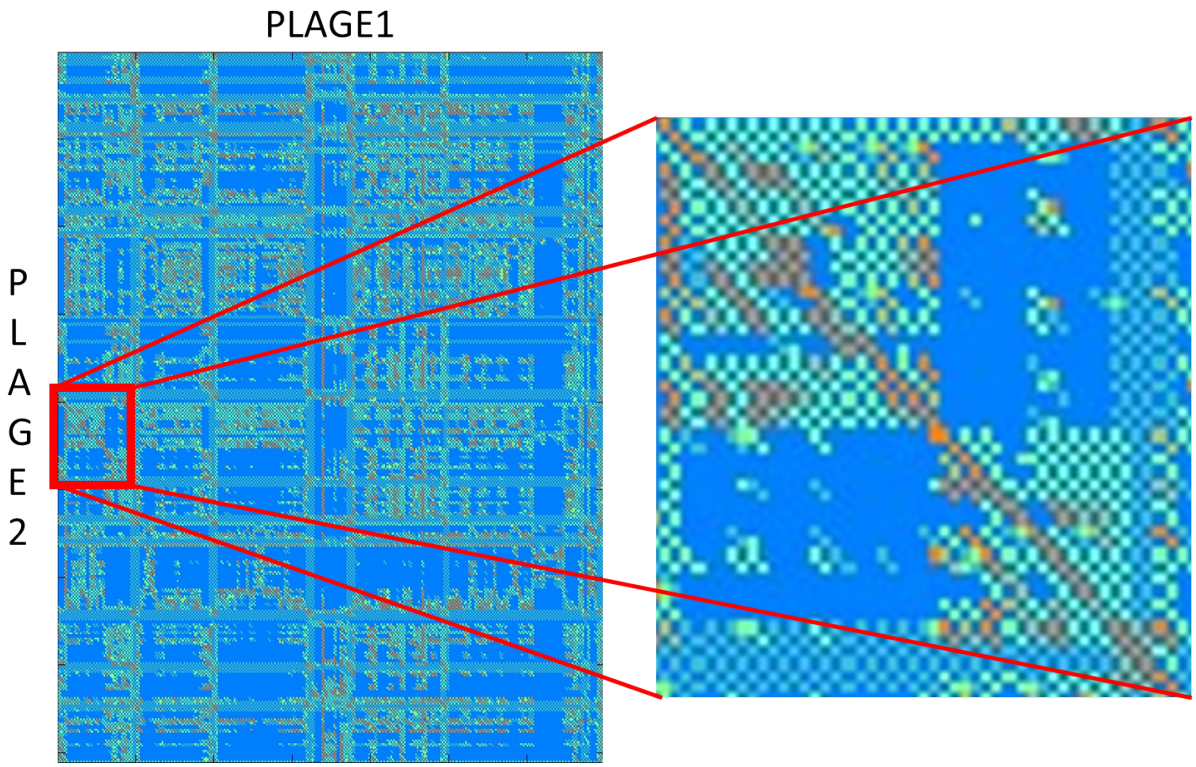



FIGURE 5.3 – Exemple de matrice de similarité sur deux plages publicitaires à partir de descripteurs sonores.

Focalisons, à titre d'exemple, sur le bloc original agrandi de la figure 5.3, à droite. Il s'agit d'un même spot diffusé sur les deux plages publicitaires : la diagonale possède des coefficients élevés car il s'agit de la comparaison d'une séquence avec elle-même.

Cette image de droite peut être vue comme une matrice d'auto-similarité du spot publicitaire. Nous pouvons nous apercevoir que ce spot peut être décomposé temporellement en deux parties : la première correspond à un monologue d'une voix off et la seconde coïncide avec le début du jingle musical. Il est donc assez logique que les deux zones soient très différentes car l'une contient de la parole et l'autre de la musique.

Ce travail a ensuite évolué de manière assez cohérente par l'association des descripteurs audio et vidéo : la fusion de matrices de similarité (une par descripteur). Suivant l'enregistrement traité, certains descripteurs sonores et/ou visuels se sont montrés particulièrement pertinents pour segmenter en zones stratégiques, représentatives du contenu ou de la structure du document.

Néanmoins, cette fusion des matrice de similarité nous a poussés à proposer une stratégie de sélection des descripteurs pertinents en vue d'obtenir la matrice de similarité globale. Ces améliorations sont décrites dans la section suivante qui se fonde sur les matrices de similarité pour « chapitrer » tout type d'enregistrement audiovisuel.

## 5.2 Chapitrage

À travers le projet ANR CHAPITRE [16], dans un contexte d'enrichissement de fonctionnalités dans la PVR (*Personal Video Recorder*) et la VoD (*Video on Demand*), nous nous sommes focalisés sur la présentation synthétique du contenu avec un accès direct aux parties pertinentes de celui-ci. En effet, ce projet RIAM [17], avec nos partenaires NPTV, Expway et NDS, avait pour objectifs d'offrir une ou plusieurs vues intelligentes d'un contenu numérique à un téléspectateur et de lui permettre d'interagir.

La notion de chapitre correspond pour nous à une unité de structure dont son contenu doit être homogène, c'est-à-dire « similaire ».

L'équipe SAMoVA s'est focalisée sur le développement de méthodes de « chapitrage » de contenu vidéo, sachant que dans l'état de l'art très peu de méthodes permettaient de faire une macro-segmentation automatique et indépendante du contenu.

Ceci était d'autant plus difficile que nous avions des contraintes de temps réel afin de fournir les méta-données correspondant au « chapitrage » en direct (ou léger différé) sur une Set Top Box (décodeur TV).

Nous nous sommes appuyés sur les matrices de similarité (principalement d'auto-similarité) vues dans la section 5.1. Celles-ci doivent néanmoins évoluer sur plusieurs points :

- permettre une fusion moins naïve des descripteurs,
- se spécialiser au contenu traité, qui peut être très divers,
- être calculées partiellement sur des fenêtres glissantes afin de limiter la complexité des algorithmes et répondre à la contrainte du pseudo temps-réel (léger différé).

### 5.2.1 Méthode de chapitrage

La première étape a consisté à proposer deux types de méthodes de classification, que nous nommons **grammaires** : l'une spécialisée, l'autre générique.

16. Classification Hiérarchique et Automatique de vidéos pour une Plate-forme Interactive de Télévision numéRique Enrichie

17. http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-05-RIAM-0016

La **grammaire spécialisée** consiste à confronter l'enregistrement à des exemples d'événements recherchés :

- pour un journal télévisé, nous utilisons un plan du présentateur,
- pour un match de foot, l'exemple peut être un but,
- pour un jeu télévisé, le générique de transition a été choisi,
- pour une série télévisuelle, un plan du héros peut être pertinent.

La **grammaire générique** ne détecte pas d'événement spécifique afin de chapitrer. Elle est fondée sur 3 niveaux qui correspondent à l'intensité des variations entre segments homogènes :

- le niveau 0 doit correspondre à la durée d'un chapitre, entre 3 et 5 minutes,
- le niveau 1 est de l'ordre de la séquence (environ une minute),
- le niveau 2 est équivalent à une scène, de 10 à 20 secondes.

Les caractéristiques sonores et visuelles utilisées sont les mêmes que précédemment pour les matrices de similarités (voir section 5.1.2), à savoir 12 descripteurs vidéo et 9 descripteurs audio. Chaque paramètre extrait, renvoie une valeur par unité de temps (à savoir une image, $1/25^{\text{ième}}$ de seconde).

Selon la nature des contenus traités, telle ou telle caractéristique peut s'avérer plus ou moins pertinente pour mettre en évidence un changement de chapitre. Nous jugeons qu'un descripteur est pertinent s'il est stable dans le temps, par rapport à l'événement traité.

Ainsi, pour identifier l'absence (ou la présence) de variations fortes de similarité dans la matrice d'un descripteur, nous calculons la dérivée :

$$D_f(i) = |f(i) - f(i+1)| \tag{5.2}$$

à partir de la fonction $f(i)$ qui correspond à la moyenne des coefficients à un instant $i$.

Ensuite, $D_f$ est quantifiée linéairement de manière à être exprimée par une valeur entière comprise entre 0 et 100. Un histogramme $h(v)$ des valeurs quantifiées est ensuite calculé.

Rappel : une caractéristique est jugée d'autant plus pertinente qu'elle est stable dans le temps. Nous cherchons donc les valeurs les plus élevées de l'histogramme pour les $p$ variations les plus faibles. Nous avons fixé arbitrairement $p$ à 10 %.

Nous calculons alors le taux de variation $w_k$ du descripteur $k$ :

$$w_k = \frac{\sum_{v=0}^{p} h(v)}{\sum_{v=0}^{100} h(v)} \tag{5.3}$$

$W_k$ est alors le poids affecté à la matrice de similarité calculée sur le descripteur $k$ :

$$W_k = \frac{w_k}{\sum_k w_k} \tag{5.4}$$

Les matrices de similarité (ou d'auto-similarité suivant la méthode utilisée) sont obtenues pour chaque caractéristique par la méthode vue précédemment (voir section 5.1.1 pour le calcul d'IQR et du taux de couverture).

Les matrices sont fusionnées en tenant compte de la pondération ainsi définie pour ne plus en former qu'une seule matrice de similarité finale (issue des 21 descripteurs audio-vidéo).

Une étape de filtrage est appliquée sur la matrice de similarité finale. Nous recalculons la fonction $f(i)$ qui rend compte de la similarité moyenne de chaque séquence à un instant $i$ d'un document avec toutes les séquences du document comparé, toute caractéristique confondue. Deux types d'information fournie par $f$ sont pertinentes à nos yeux pour la suite :

- les fortes variations ponctuelles et éparses correspondent à un changement de chapitre,
- les zones longues où la similarité est stable correspondent à des zones homogènes.

Nous appliquons un filtre coupe-bande (type filtre médian) pour obtenir la fonction filtrée $F$. Comme précédemment, nous calculons la dérivée temporelle $D_F$.

Un changement de chapitre doit correspondre à une variation particulièrement remarquable de la similarité moyenne (i.e. une valeur élevée de $D_F$). Nous utilisons ici la notion de durée moyenne de segment à détecter, lorsque celle-ci est connue *a priori* (cf. grammaire générique). Soit $d$ cette durée moyenne. Nous calculons l'Importance Relative $IR(i)$ d'une variation $D_F(i)$ dans une fenêtre définie par $d$ autour de l'instant $i$ par :

$$IR(i) = \frac{D_F(i)}{\max\limits_{j=i-d/2}^{i+d/2} D_F(j)} \tag{5.5}$$

Une limite de segments (changement de chapitre) est détectée quand : $IR(i) >$ seuil. En pratique, seuil $= 1$.

Vis-à-vis des deux grammaires utilisées, le système se décline en deux sous-systèmes (voir figure 5.4) : générique et spécialisé.

Dans le **mode générique**, à partir des caractéristiques sonores et visuelles extraites, le flux est comparé avec lui-même. Ce mode consiste à construire une *matrice d'auto-similarité partielle* à partir des 21 descripteurs. La matrice est partielle dans le but de traiter des flux.

En effet, il n'est donc plus possible d'attendre la fin de l'extraction des caractéristiques pour lancer le calcul des coefficients de la matrice. Pour palier ce problème, nous avons décidé de tronçonner le flux en « buffers » d'environ 5 minutes chacun. Chaque « buffer » est composé de la dernière moitié des valeurs obtenues sur le « buffer » précédent, et de la première moitié des valeurs du « buffer » suivant.

Lorsque toutes les valeurs des caractéristiques sont obtenues pour un « buffer », la sous-partie de la matrice correspondante peut être calculée, et la détection des ruptures permet ensuite le « chapitrage ». Ceci induit une latence mais le temps réel peut être garanti.

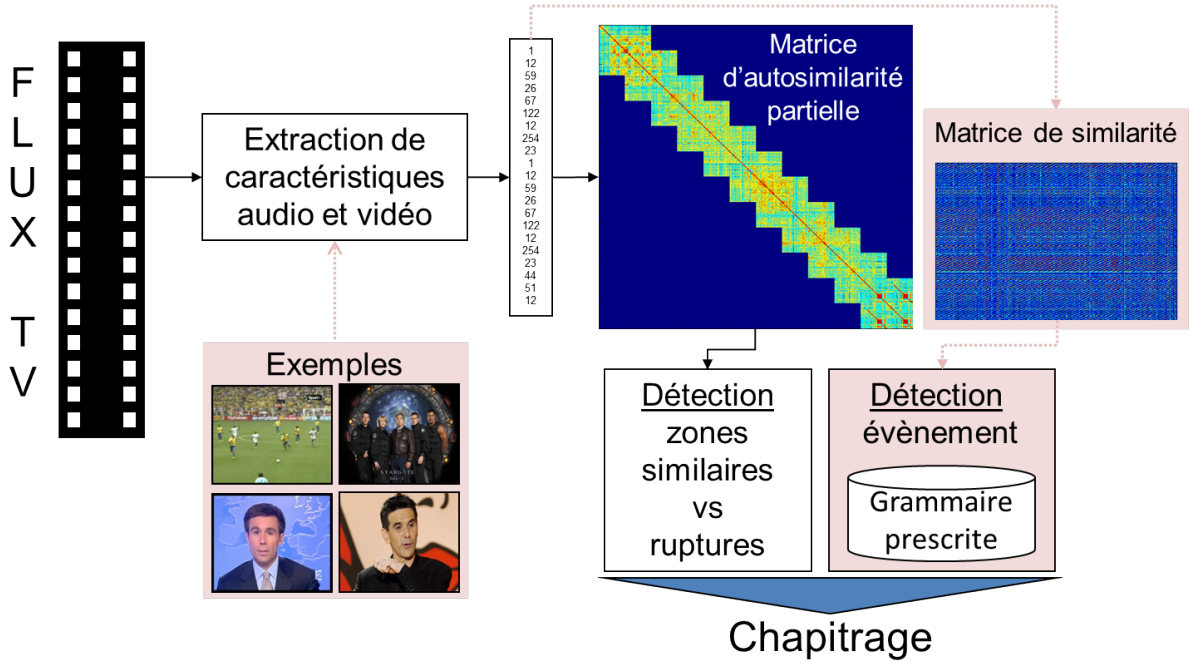



FIGURE 5.4 – Système de « chapitrage » par les approches générale et spécialisée (en rose).

Dans le **mode spécialisé**, toujours à partir des caractéristiques sonores et visuelles extraites, le flux est comparé à un enregistrement audiovisuel court (exemple) présentant les événements structurant du contenu à traiter. La matrice de similarité est alors calculée dans son intégralité. Une variation élevée de $F(i)$ correspond à un passage d'un événement structurant à un autre type de contenu.

### 5.2.2 Résultats de « chapitrage »

Voici quelques faits marquant de cette méthode de « chapitrage ».

L'approche spécialisée (en utilisant des exemples et la matrice de similarité) appliquée aux journaux télévisés de *France 2* (avec un plan du présentateur comme exemple) a permis de segmenter l'émission en plateaux/reportages de manière parfaite.

Cette même approche sur des *matchs de la coupe du monde de football de 2006* (confrontés à 7 exemples de but provenant du championnat de France), permet de détecter pratiquement tous les buts et quelques actions « chaudes » (pénaltys, coup francs ou cartons). Ceci s'explique car les actions chaudes ont des caractéristiques assez similaires aux buts, à savoir un volume sonore qui augmente et des rediffusions sous forme de ralentis.

La figure 5.5 illustre la détection de deux zones ponctuelles de fortes similarités (buts) et une zone complètement différente du reste (la publicité).

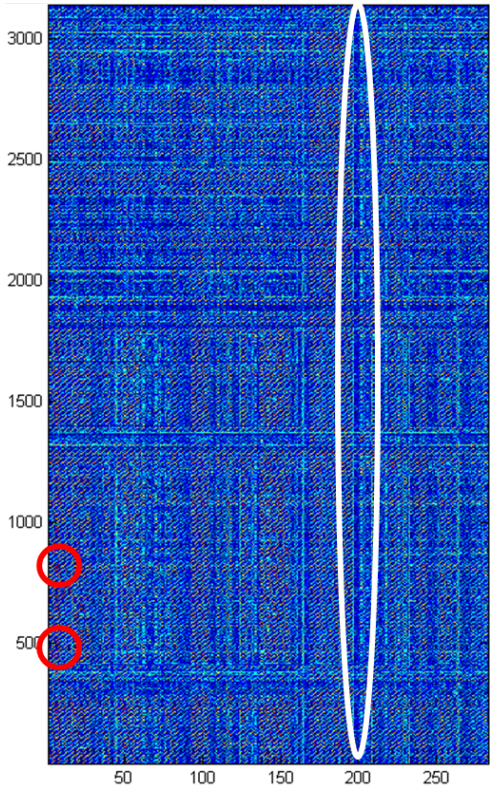



FIGURE 5.5 – Exemple de matrice de similarité d'un match de football : les ronds rouges correspondent aux buts et l'ellipse blanche à la publicité.

L'approche générique (en utilisant la matrice d'auto-similarité) appliquée à la série *Stargate* obtient des résultats qui sont, de manière générale, concordants avec les chapitres des DVD fournis : 25 % des chapitres sont même identiques.

Cette même approche a été appliquée à un jeu télévisé *les z'amours* qui a la particularité de contenir différentes phases de jeu (8) : générique de début, présentation des couples, partie 1 (question pour les hommes puis réponses des femmes), partie 2 (question pour les femmes puis réponses des hommes), partie 3 (finale), générique de fin.

Les résultats obtenus sont encourageants : les 8 phases sont retrouvées ! Par contre, des fausses alarmes sont présentes, notamment à cause des applaudissements, des rires, de la musique et des animations visuelles.

À partir de la matrice d'auto-similarité de la figure 5.6, il est intéressant de remarquer une zone (repérée par un cercle noir) composée de fortes valeurs (2 rectangles rouges symétriques). Il s'agit de la présentation sous le même fond musical des cadeaux à gagner par les deux couples de candidats.

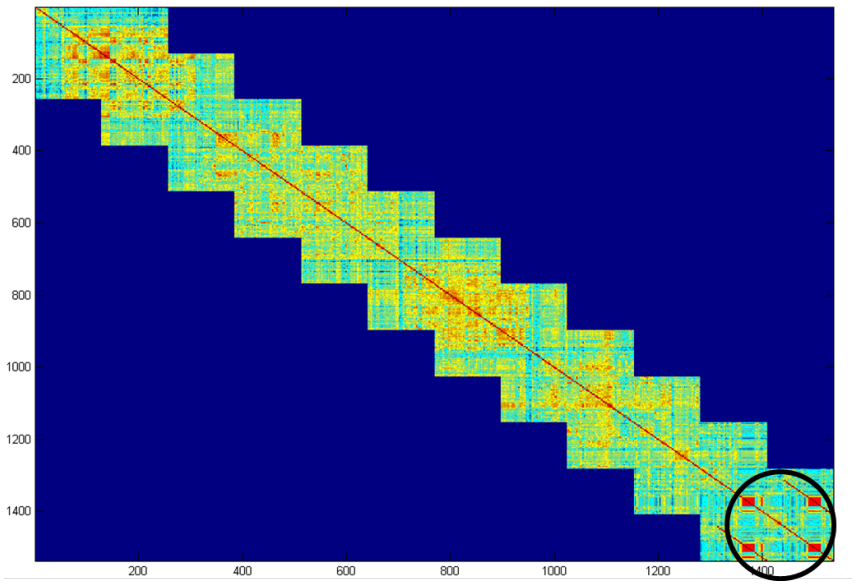



FIGURE 5.6 – Exemple de matrice d'auto-similarité obtenue sur un jeu télévisé.

Ce travail de « chapitrage » ainsi que l'étude précédente sur la similarité sont directement fondées sur les données extraites du signal. Or, celles-ci ne sont pas directement assimilables à des concepts audiovisuels. Dans cette dernière étude sur l'organisation de contenus audiovisuels, nous tentons de résorber ce *fossé sémantique* (« semantic gap »). Nous essayons de réduire la distance qui existe entre un concept audiovisuel, pouvant être exprimé et compris par un humain, et sa représentation numérique, directement interprétable par la machine.

## 5.3 Organisation de contenus audiovisuels

Comme les travaux précédents, nous reprenons la notion de similarité entre contenus audiovisuels. Or, cette notion est purement subjective. Une personne peut dire que deux documentaires se ressemblent parce que les narrateurs ont le même accent, parce que les couleurs du plateau sont sensiblement les mêmes, ou encore parce que les sujets traités sont voisins. Afin de respecter cette vision subjective de la similarité, nous avons créé un système qui puisse déterminer quelles caractéristiques auditives et/ou visuelles sont à mettre en relation pour pouvoir en dégager un sens qui corresponde à une tâche organisationnelle suggérée par l'utilisateur.

Voici un exemple de scénario d'utilisation permettant d'organiser une base de données audiovisuelles en fonction de la signalétique du programme « -10 », « -12 », « -16 », etc.

Nous positionnons (ancrons), à l'aide d'une interface dédiée, quelques documents de manière à rendre visuellement compte des écarts de violence du contenu.

Ensuite, nous demandons au système d'inférer notre démarche sur la totalité de la base. Nous disposons d'une batterie de caractéristiques extraites de ces vidéos, dont par exemple la quantité de mouvement et l'énergie sonore. Le système trouve pertinent d'utiliser ces descripteurs, et

interprète la tâche organisationnelle comme une volonté de répartir sur l'interface, de manière homogène, les entités documentaires selon ces critères.

Il établit alors une relation entre la position des documents d'une part et les valeurs descriptives correspondant au couple (quantité de mouvement, énergie) d'autre part, pour rapatrier et organiser le reste de la base.

### 5.3.1 Système d'organisation de contenus audiovisuels

La figure 5.7 résume les trois principales phases du processus de création de la mesure de similarités : « apprentissage », « prédiction » et « visualisation ».

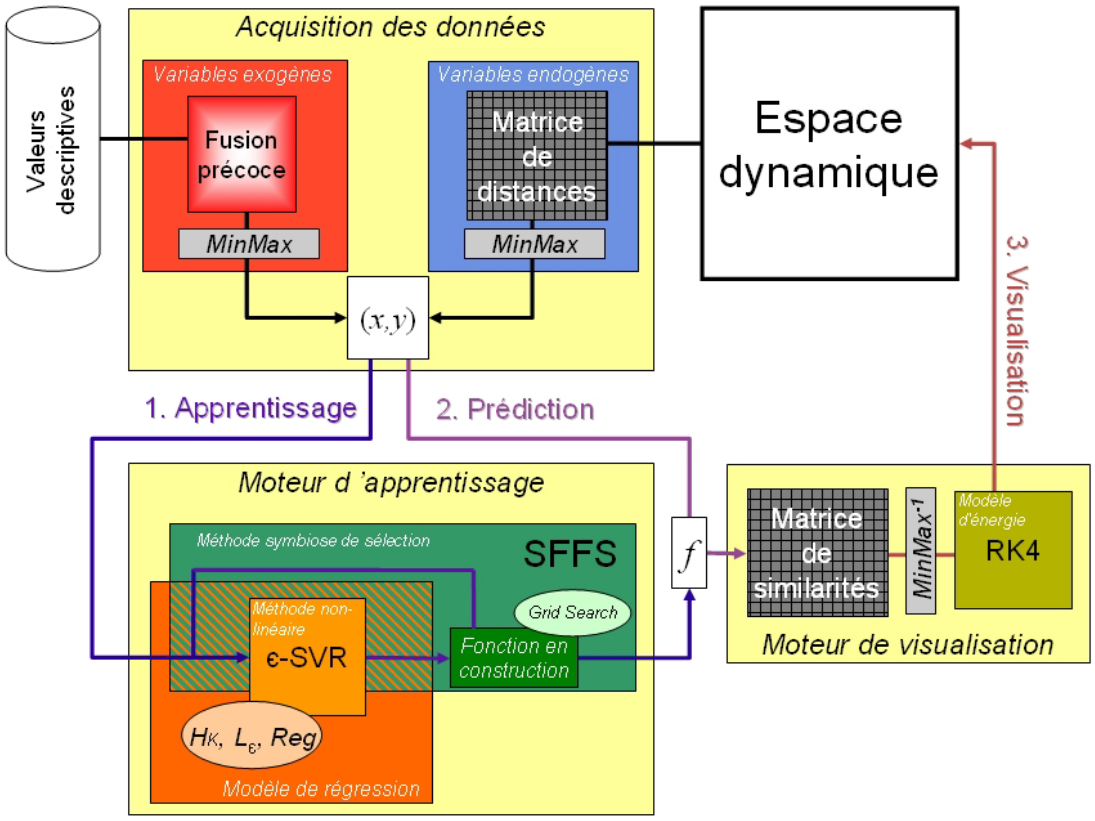



FIGURE 5.7 – Schéma général du système d'organisation de contenus audiovisuels.

Considérons un utilisateur qui positionne quelques contenus sur l'espace dynamique et ancre certains d'entre eux. Il lance la procédure d'apprentissage, en demandant ou non le rapatriement d'un ou plusieurs contenus (choisis ou laissés à la discrétion du système).

L'**apprentissage** est le cœur de ce travail. Il se déroule en trois phases :

1. la création des variables endogènes,
2. la création des variables exogènes,
3. l'apprentissage de la fonction de régression.

Les **variables endogènes** correspondent aux variables à prédire, c'est-à-dire aux distances entre les documents que l'utilisateur a positionnés sur l'interface. Pour $M$ documents, nous avons donc $N$ valeurs de distance :

$$N = \frac{M(M-1)}{2} \tag{5.6}$$

Une normalisation $MinMax$ permet de ramener chaque distance $d_{i,j}$, entre le document $i$ et le document $j$, sur l'intervalle $[0,1]$, notée $\hat{d_{i,j}}$.

Ces distances $\hat{d_{ij}}$ sont rangées dans un vecteur endogène $v_{en}$ de dimension $N$ :

$$v_{en} = (\hat{d_{12}}, \hat{d_{13}}, ..., \hat{dij}, ..., \hat{d_{(M-2)M}}, \hat{d_{(M-1)M}}) \tag{5.7}$$

Les **variables exogènes** sont quant à elles les variables prédictives, c'est-à-dire les valeurs des descripteurs audiovisuels calculés. Les variables utilisées sont les mêmes que pour les matrices de similarité et le « chapitrage » : 12 descripteurs visuels et les 9 descripteurs sonores (voir section 5.1.2 pour plus de détails). Les valeurs descriptives utilisées sont les moyenne, variance, minimum et maximum de ces descripteurs « bas niveau ». Celles-ci sont combinées (fusion précoce) et nous obtenons ainsi un vecteur $v_{ex}$ de 84 valeurs. Ce vecteur est également normalisé pour avoir des valeurs entre 0 et 1.

Les variables exogènes $x_i$ $i \in 1, ..., N$ correspondent aux $N$ concaténations de deux vecteurs normalisés, issus des $M$ contenus du corpus d'apprentissage, en respectant l'ordonnancement des variables endogènes.

Le moteur d'**apprentissage** utilise une régression $\epsilon$-SVR univariée, fondée sur la concaténation de valeurs descriptives sélectionnées par un algorithme de type SFFS. Ses hyper-paramètres sont optimisés grâce à une méthode de recherche partielle par grille.

De nombreuses méthodes existent dans la littérature afin de concevoir des mesures de similarité. Citons par exemple, la *Principal Component Analysis* (PCA, [Pearson 01]), le *Multi-Dimensional Scaling* (MDS, [Cox 94]), la *Relevant Component Analysis* (RCA, [Bar-Hillel 05]), la *Discriminative Component Analysis* (DCA, [Hoi 06]), etc. Cependant, elles ne sont pas en accord avec notre philosophie : ces distances modifient la nature quantitative de la relation de similarité dans l'espace de description, pour mieux la qualifier dans l'espace de représentation. Au lieu de chercher à concilier la similarité entre les deux espaces, elle la force.

Nous nous sommes donc tournés vers la prédiction statistique. En effet, nous interprétons une relation de similarité numérique entre les contenus comme un phénomène, au sens probabiliste du terme, dont il est possible d'estimer le comportement à partir d'un jeu d'observations.

Dans notre système, définir une régression revient à créer une relation qui s'appuie sur un ensemble de valeurs observées à partir d'un ensemble de variables exogènes pour estimer des variables endogènes. Il se peut toutefois que le type de dépendance entre les données, qui caractérise l'expression de la similarité, ne soit pas linéarisable, ou ne soit pas connu *a priori*. Dans ce cas, une fonction non-linéaire est utilisée afin de prédire la similarité entre les données. Notre

volonté d'utiliser toute forme de donnée numérique sans conserver d'information sur leur comportement ne nous permet pas d'anticiper sur le type de dépendances qui pourrait exister entre l'agencement des contenus et leurs descriptions (autrement formulé, entre variables endogènes et exogènes). Nous devons accorder un maximum de flexibilité à notre espace d'hypothèses, et emploierons pour ce faire une fonction non-linéaire afin de piloter notre modèle.

De nombreuses méthodes sont à disposition : *Non-Linear Least Square* (NLLS, [Kelley 99]), $\epsilon$ *Support Vector Regression* ($\sigma$-SVR, [Cortes 95]), etc.

Parmi elles, nous avons choisi d'utiliser la **régression par machines à vecteur de support** ($\epsilon$-SVR), du fait de ses nombreux avantages :

- rapidité de l'apprentissage pour assurer une organisation en « temps interactif » (délai inférieur à 10 secondes) sur l'interface de visualisation,
- méthode parcimonieuse qui permet de gérer l'imprécision potentielle apportée par l'utilisateur lors de l'agencement des objets sur l'espace dynamique (seules les distances réellement significatives interviendront dans le modèle, permettant ainsi de potentiellement réduire une partie du bruit),
- implémentation disponible via des bibliothèques telle *lib-svm*[18].

Au-delà du choix de la fonction utilisée, la performance d'un modèle de régression est très souvent fonction de la qualité des données d'apprentissage. Pour avoir la prétention d'expliquer un phénomène, il faut que les valeurs descriptives (les témoins de ce phénomène) soient le moins bruitées possibles. Ce lien entre les données et le modèle se fait via des algorithmes de réduction de la dimensionnalité.

Là encore, de nombreuses méthodes existent que nous pouvons classer en deux catégories : les méthodes d'extraction de caractéristiques et les méthodes de sélection de paramètres.

Les méthodes d'extraction de caractéristiques peuvent être linéaires : soit non-supervisée comme l'*Independant Component Analysis*, (ICA, [Comon 94]), soit supervisée comme la *Linear Discriminant Analysis* (LDA, [Haeb-Umbach 92]). Des méthodes non-linéaires existent également en mode non-supervisée (telle *Isomap*, [Tenenbaum 00]) et supervisée (comme S-Isomap, [Geng 05]).

Les méthodes de sélection de paramètres peuvent être optimales avec solution unique (telle *Branch and Bound*(BB) [Narendra 77]), soit à solution multiple (comme l'algorithme Las Vegas (LVS), [Brassard 96]). Ces méthodes peuvent également être sous-optimales : les plus connus étant les algorithmes séquentiels à solution unique *Sequential Forward Selection* (SFS) et *Sequential Backward Selection* (SBS), [Devijver 82]. Le *Beam search* [Aha 95] permet la conservation d'un ensemble de solutions.

18. http://www.csie.ntu.edu.tw/cjlin/libsvm/

Nous avons opté pour l'**algorithme séquentiel sous-optimal SFFS** (*Sequential Floating Search methods*) car :

- il détermine de lui-même le nombre de passes avant et arrière à effectuer,
- l'échafaudage itératif de la variable exogène par l'ajout ou la suppression de valeurs descriptives permet de théoriquement garder à l'écart des descriptions qui viendraient perturber le modèle de régression,
- la fonction d'évaluation (la mesure de performances liée à la méthode d'apprentissage) est l'erreur quadratique moyenne,
- la méthode d'évaluation est celle utilisée pour déterminer les hyper-paramètres du modèle (validation croisée),
- son critère d'arrêt est également l'erreur quadratique moyenne : s'il n'y a plus d'amélioration au sens du critère de performances du modèle, l'algorithme se termine.

La deuxième partie du schéma 5.7 concerne la **prédiction**, c'est-à-dire le rapatriement d'autres documents de la base dans l'interface en tenant compte de l'organisation de l'utilisateur déjà présente dans l'interface. Ceci est effectué à partir de la fonction de régression en renvoyant les éléments qui sont sensés le plus perturber l'organisation courante : il s'agit d'une maximisation de la similarité moyenne entre un document à rapatrier et tous les documents déjà présents.

La dernière partie de la figure 5.7 correspond au **moteur de visualisation**. Une fois le rapatriement effectif, il reste à transcrire ce dernier sous la forme d'une matrice de distances pour qu'il soit visualisable sur l'espace dynamique (interface).

Dans un premier temps, les échelles des similarités sont changées par une normalisation $MinMax$ inverse afin de faire coïncider le maximum (resp. minimum) des valeurs estimées avec la valeur maximale (resp. minimale) des valeurs de distances du corpus d'apprentissage.

Pour obtenir une matrice de distance visualisable, il reste à garantir l'inégalité triangulaire. La résolution de ce problème est laissée à un modèle d'énergie. Son principe est simple : une matrice de distances est calculée à partir des positions initiales des sommets du graphe. Cette matrice est modifiée à intervalles de temps réguliers dans le but de se rapprocher le plus possible de la matrice de similarités, sous la contrainte des différentes forces agissant entre les sommets : force d'attraction, force de répulsion, viscosité et vitesse.

Pour calculer les nouvelles positions des sommets du graphe, nous avons choisi d'utiliser l'algorithme de Runge-Kutta à l'ordre 4 (RK4), couramment employé dans le domaine de la physique. Notons que l'interface visuelle a été développée en OpenGL [19] grâce à la bibliothèque graphique Clutter [20].

19. http://www.opengl.org
20. http://www.clutter-project.org

### 5.3.2 Interface du système d'organisation de contenus audiovisuels

La figure 5.8 présente une capture d'écran annotée de notre interface. Elle se compose de quatre zones principales :

- la zone I est une fenêtre dans laquelle il est possible d'organiser des contenus audiovisuels,
- la zone II fournit des informations sur les contenus de la zone I,
- la zone III présente une vue horizontale du reste de la base de données. Un curseur de défilement horizontal permet de s'y déplacer, d'y sélectionner des contenus qui peuvent être rapatriés vers la zone I afin d'y être organisés,
- la zone IV propose des outils d'interaction pour agir sur l'ensemble de la GUI : sauvegarde et chargement des données, quitter, lancement de l'apprentissage via le rapatriement d'un ou plusieurs contenus, etc.


FIGURE 5.8 – Capture d'écran du système d'organisation de contenus audiovisuels.

La zone I, appelé espace dynamique, est une interface en deux dimensions de type WIMP (Windows, Icons, Menus, Pointing device) [van Dam 97]. Une fenêtre de visualisation (partie ①) permet d'appréhender des contenus représentés sous forme d'icônes (partie ③). Nous pouvons les manipuler (sélectionner, déplacer, consulter) à l'aide de la souris (partie ②). Une vignette représentative du contenu audiovisuel constitue le corps de l'icône : des fonctionnalités permettant d'entrer en interaction avec le système sont présentées dans un menu (partie ④).

Différentes expériences ont été réalisées afin de valider les différents choix techniques et théoriques : le choix des paramètres audio-vidéo « bas-niveau » la durée d'apprentissage (au pire de 10 secondes), la méthode de régression ($\epsilon$-SVR) et la méthode de sélection de paramètres (SFFS) avec optimisation partielle. Les résultats obtenus sur nos différentes expériences sont encourageants.

Nous avons montré que notre système propose une aide non négligeable pour organiser des contenus audiovisuels qui possèdent certaines spécificités propres à leur signal. Par exemple :

- le classement de 96 notes de musique à partir de l'organisation de 7 notes de départ dans l'interface ne donne aucune erreur,
- l'organisation de ces 96 notes en instruments suivant les 3 niveaux de la taxonomie de Peeters [Peeters 03] est parfaite à partir de 17 éléments en apprentissage (80 % de classification correcte avec 7 éléments),
- la segmentation plateaux / reportages de 43 extraits d'émissions du « Morning Café » est également quasi parfaite (1 erreur seulement) à partir de 7 éléments en apprentissage,
- l'organisation de 100 contenus de l'INA (306 minutes extraites de « Cinq colonnes à la une », « Les actualités françaises », « La minute de Monsieur Cyclopède », « Reflets de Canne » et « Spécial Sport ») a été effectuée en 21 minutes, soit environ quinze fois moins de temps qu'il n'en faudrait pour visionner l'intégralité du corpus.

De plus, nous avons testé notre prototype avec des utilisateurs « grand public ». Les fonctionnalités propres à l'espace de travail ont été particulièrement appréciées.

# Discussion

Nous venons d'aborder dans ce chapitre, la segmentation audiovisuelle à travers deux angles de vue : les segmentations fondées sur l'intervenant et les segmentations fondées sur la similarité. Le cœur de la recherche effectuée se situe principalement sur le couplage entre l'audio et la vidéo à partir de paramètres sonores et visuels « bas-niveau ». Voici quelques conclusions et perspectives directes de chacun de ces travaux.

## Segmentation autour de l'intervenant

Cette partie regroupe quatre travaux qui, partant de la détection des intervenants et de leur caractérisation, nous amènent vers une structuration aussi bien de contenus formatés (nous parlerons alors de programmes) que de vidéos personnelles (nous utiliserons alors le mot activités).

Grâce à nos acquis sur la segmentation sonore, nous avons naturellement combiné notre méthode de segmentation en locuteurs avec une méthode de détection de costumes (et de visages) afin de segmenter des contenus en intervenants. L'audio et la vidéo étant parfois asynchrones, nous avons proposé une méthode de fusion atypique par matrice de co-occurrence.

L'évolution et l'amélioration de ce travail sont principalement assujetties à l'amélioration des techniques de segmentations sonores et visuelles car notre fusion donne des résultats quasi-parfaits sur des segmentations manuelles.

Les intervenants ont ensuite été caractérisés à travers une classification IN, OUT ou OFF inédite. Celle-ci permet de distinguer si la personne qui parle est aussi à l'écran (IN), a été ou sera à l'écran (OUT) ou n'est jamais visible (OFF). Ce travail, fondé sur un automate, pourrait évoluer de diverses façons.

Tout d'abord sur le plan des descripteurs, nous sommes en train de travailler à l'aide de la caméra Kinect [21] développée par Microsoft pour le projet ANR RIDDLE [22] sur des paramètres visuels. Nous devons notamment connaître l'intentionnalité d'une personne à vouloir communiquer avec un robot. Dans ce contexte, le robot équipé de la Kinect, analyse à la fois la voix de la personne et ses mouvements (corps, buste, tête et lèvres) pour savoir si la personne a besoin de lui. L'utilisation des « random forest » à partir des canaux couleurs et intensités de la Kinect, dans la lignée des travaux de Fanelli [Fanelli 13] semblent très prometteurs.

21. Microsoft Corp. Redmond WA. Kinect for Xbox 360
22. http://projects.laas.fr/riddle/

Ensuite, sur le plan du classifieur, il semble intéressant de prendre en compte l'aspect temporel : un HMM semble alors tout à fait pertinent. Nous pourrons, par exemple, nous appuyer sur l'expérience acquise lors du projet IMMED sur ce type de classifieur et sur l'utilisation de fusion à différents niveaux. Une fusion précoce, par concaténation de nos différents paramètres, serait alors privilégiée.

Les intervenants, par l'intermédiaire de leurs rôles et de leurs interactions, ont par la suite été au cœur d'une structuration à deux niveaux. La première, plutôt « macro », permet de se rapprocher d'une grille de programme. La seconde plutôt « micro », met en évidence les zones conversationnelles d'un enregistrement.

Ce travail, bien qu'appliqué à des contenus audiovisuels, souffre de son approche mono-média. J'ai d'ailleurs hésité à le mettre dans le chapitre 1 de ce manuscrit avec les segmentations sonores. Néanmoins, il m'a semblé plus judicieux de le placer dans ce chapitre car ce travail exploite des segmentations sonores « bas-niveau » dans le but de la structuration audiovisuelle.

Afin de faire évoluer notre système, il serait pertinent d'exploiter l'information disponible dans le flux audiovisuel, par exemple en utilisant :

- une segmentation en intervenants. Elie El Khoury a montré dans sa thèse que la SRL est améliorée par l'utilisation de visages parlants [El Khoury 10].
- le contexte visuel dans lequel apparaît certains rôles (plateau pour le présentateur, reportage pour le journaliste, etc.).

Les programmes représentent l'unité des enregistrements radiophoniques et télévisuels : ils correspondent aux émissions ou parties d'émission. Lorsque nous traitons des vidéos personnelles cette unité existe toujours : nous pouvons d'ailleurs la nommer « programme de la journée ». Nous avons préféré utiliser le terme « activité ».

L'étude de la segmentation en activités du quotidien s'est avérée très difficile pour différentes raisons :

- le nombre d'activités à détecter est très important,
- le dispositif d'acquisition audio-vidéo a une qualité inférieure aux enregistrements audiovisuels traités jusque là. Ceci est particulièrement vrai sur le son pour lequel la fréquence d'échantillonnage n'est que de 12 kHz.
- le dispositif étant porté par la personne sur son épaule, des problèmes se posent : le mouvement important, les frottements éventuels et la voix plutôt forte de la personne par rapport à son environnement.
- chaque enregistrement correspond à un nouveau contexte sonore et visuel : nouvelle personne, nouveau lieu, nouveaux objets manipulés, etc.

Malgré cela, les résultats obtenus avec un HMM hiérarchique et une fusion intermédiaire sont corrects : 44,2 % d'accuracy. La partie applicative en aval de cette étude est même très prometteuse.

En effet, la partie clinique :

- a validé notre outil : celui-ci étant capable de restituer des éléments objectifs, clairs et utilisables concernant la capacité de réalisation des activités de la vie quotidienne de patients potentiellement atteints de démence,
- a confirmé que l'interface de navigation couplée à l'analyse automatique permet de ne visionner que les seuls moments clés utiles au professionnel de santé,
- a montré l'intérêt pour les professionnels de santé d'une telle approche dans le contexte des pathologies démentielles. Cette étude permet une appréhension nouvelle d'un élément du diagnostic qui est le retentissement à domicile des troubles précoces.

Des trois modalités (dynamique, statique et sonore), l'audio est pour la majorité des activités le plus pertinent. La mise en place d'autres descripteurs de bruit me semble alors indispensable. Par l'intermédiaire du projet RIDDLE, cité précédemment, la reconnaissance d'activités sera étudiée. Dans ce contexte, nous comptons proposer des paramètres spécifiques de la manipulation d'objets. Une approche fondée sur les matrices non-négatives, comme le propose Heittola [Heittola 11], permettrait peut être d'effectuer une séparation de sources au préalable de notre classification par HMM. Dans son étude, en fixant le nombre de sources à 4, il améliore les résultats de détection de 61 classes sonores.

Dans cette partie, la segmentation audiovisuelle, en plus d'être ciblée sur l'intervenant, est vue de manière plutôt linéaire. Dans la seconde partie, nous utilisons une approche plus globale afin de segmenter les enregistrements et ainsi retrouver l'organisation d'un document audiovisuel et le caractériser.

## Segmentation autour de la similarité

Dans cette partie, trois études traitant de la similarité intra ou inter-documents ont été proposées. Ces études avaient toutes le même objectif : structurer et organiser les documents audiovisuels.

Ma recherche a débuté sur ce sujet par la lecture de la thèse de Siba Haidar [Haidar 05], encadrée par Philippe Joly. Dans celle-ci, des matrices de similarité avait été proposées afin de comparer des documents. Ces matrices étaient obtenues à partir de descripteurs visuels uniquement. Nous avons alors adapté la méthode afin qu'elle puisse également fonctionner sur des critères sonores. Nous avons ainsi proposé et testé des descripteurs audio.

Les résultats obtenus furent probants. Comme pour la vidéo, une mise en valeur de la similarité apparaît sur la matrice : celle-ci est soit identique, soit complémentaire à celle de la vidéo. À partir de la matrice, il est alors possible de distinguer la structure d'un document.

Les principales limites étaient le choix des paramètres audiovisuels pertinents et l'automatisation de la structuration. Celles-ci ont été résolues grâce au « chapitrage ».

L'étude du chapitrage s'est effectuée suivant deux modes :

- le mode générique effectue un calcul de matrice d'autosimilarité partielle à partir de l'ensemble des descripteurs extraits du flux audiovisuel. La détection de ruptures permet alors de chapitrer le contenu traité.
- le mode spécialisé compare le flux audiovisuel à des exemples d'événements, via une matrice de similarité calculée sur les descripteurs. Nous obtenons alors une structuration en événements.

L'intérêt de ce travail est de segmenter n'importe quel contenu en des unités homogènes, sans connaissance *a priori* dans le mode générique. Philippe Joly et moi-même étions très fiers de ce travail qui semblait avoir de fortes retombées possibles, notamment pour de la télévision interactive. Nous avons donc décidé, via l'université Paul Sabatier, de déposer un brevet, d'autant plus qu'une société s'était montrée intéressée. Malheureusement, pour des raisons administratives, ce dépôt a échoué et nous nous sommes finalement rabattus deux ans plus tard sur un dépôt logiciel... Ceci explique qu'aucune publication n'est été effectuée sur cette méthode.

L'évolution de ce travail est potentiellement importante. En effet, notre stratégie de sélection de descripteurs étant assez pertinente, il serait judicieux d'ajouter d'autres descripteurs sonores ou visuels. Mais, je pense principalement à des descripteurs textuels car ils peuvent être nombreux et très informatifs. Par exemple, les informations du guide des programmes, le résultat d'une transcription automatique de la parole, les sous-titrages, ou encore les textes incrustés dans la vidéo. Cependant, ils ne pourront pas tous être utilisés tels quel car il faut les mettre sur la même échelle que les paramètres actuels, c'est-à-dire à la cadence de l'image.

Le dernier travail était, sans nul doute, le plus prospectif : organiser une collection de documents par une similarité utilisateur, sachant que celle-ci doit être captée via quelques exemples de positionnement. Non content de cela, le système doit fonctionner en temps interactif afin d'être acceptable par l'utilisateur. Cette étude étant réalisée dans le cadre d'une thèse CIFRE avec l'INA, elle a été expérimentée sur des extraits vidéos provenant de l'offre Internet « Archives pour tous ».

À l'aide d'un modèle de régression $\epsilon$-SVR, nous avons fait coïncider la distance entre les documents positionnés par l'utilisateur dans l'interface (variables endogènes) et la distance entre les descripteurs audiovisuels calculés (variables exogènes). Ceci est rendu interactif grâce à un algorithme séquentiel sous-optimal SFFS qui effectue une sélection de paramètres. Là encore, ce type de thèse a nécessité de créer une interface conviviale (via OpenGL) et efficace (via l'algorithme Runge-Kutta d'ordre 4).

Les résultats sont difficiles à juger car une tâche de similarité est plutôt subjective, chaque utilisateur jugeant suivant ses propres critères. Néanmoins, en utilisant notre système pour des tâches de détection (plateau/reportage, notes de musique, instruments), les résultats furent du niveau de l'état de l'art. De plus, dans une tâche d'organisation de 100 documents inconnus, l'efficacité du système fut prouvée.

Comme pour l'étude précédente, le nombre de descripteurs pourrait évoluer : il serait intéressant là aussi d'utiliser la composante textuelle. Une autre perspective serait d'offrir une organisation multi-niveaux ou inter-niveaux c'est-à-dire organiser dans un premier temps des documents entre eux, puis dans un second temps organiser les parties d'un document entre elles (approche multi-échelles) et des parties d'un document avec des documents entiers (approche inter-niveaux). Ceci ne sera possible que si nous travaillons sur des paramètres pouvant être calculés (et comparés !) aussi bien de manière globale, sur le document en entier, que de manière locale, sur n'importe quelle partie du document.

# Conclusion et perspectives

À travers ces différentes études, j'ai tenté de montrer l'intérêt des paramètres sonores « bas-niveau » d'une part et de la combinaison de l'audio et de la vidéo d'autre part. Tout ceci dans un même cadre, la segmentation !

Après avoir donné une vision un peu caricaturale du traitement sonore, je propose quelques pistes qui me tiennent à cœur pour la suite de ma carrière. La première concerne l'environnement sonore que j'essaye de définir tant bien que mal depuis le début de mes recherches sans toutefois aller plus loin que proposer des segmentations de quelques unes de ses composantes. La seconde perspective est une généralisation de la notion d'environnement jusqu'ici sonore pour l'étendre à la vision. Il s'agit alors de caractériser le contexte sonore et visuel des enregistrements.

## Vision personnelle du traitement audio

Bien évidemment, les reconnaissances en parole, en musique et en bruit ne sont pas au même niveau d'avancement.

Le traitement automatique de la parole, fort de son expérience de plusieurs dizaines d'années, semble plus mature. Néanmoins, depuis quelques temps les idées ont du mal à se renouveler : les paramétrisations n'évoluent guère avec les MFCC qui ont toujours la part belle. Les modélisations restent assez classiques : les fusions s'empilent les unes sur les autres mais les approches de base restent les mêmes avec les HMM, GMM, SVM et réseaux de neurones comme fers de lance.

Côté musique, le traitement de contenus autres que des enregistrements en studio a fait apparaître de nouvelles problématiques :

- la segmentation, afin d'isoler une zone de musique (début et fin) plutôt que de traiter un morceau pré-découpé,
- la superposition, afin de gérer des contenus musicaux potentiellement mélangés à de la parole ou du bruit.

Le rapprochement assez récent des communautés parole et musique devrait permettre un enrichissement mutuel. Par exemple, lors de la tâche 6.5 de structuration de contenus musicaux du projet QUAERO, l'approche dite « parole » a permis une annotation en unités répétitives homogènes et l'approche dite « musique » a consisté à segmenter en éléments musicaux.

Au final, les deux visions donnèrent une annotation quasi-identique !

L'analyse de la composante bruit est sans nul doute celle qui va évoluer le plus dans les années à venir. En effet, grâce aux avancées en traitement du signal d'une part et perception acoustique d'autre part, il semble désormais envisageable de décrire les paysages sonores qui nous entourent, tels que Schafer les avait formalisés [Schafer 77].

Il distingue trois catégories sonores dans les paysages ou *soundscapes* :
- les sonorités maîtresses ou toniques, **keynote sounds**, qui jouent le rôle de fond. Le terme *key* n'est pas anodin : il signifie que c'est le son fondamental. Notre perception du paysage est conditionnée par ces sons, même si souvent ils passent inaperçus (nous nous en rendons couramment compte lorsqu'ils s'arrêtent...). Par exemple, le bruit de l'eau pour des habitants d'une maison se situant à côté d'une rivière ne s'entend plus au bout d'un certain temps.
- les sons à valeur signalétique, **signal sounds**, qui apparaissent comme des événements pour les personnes qui les entendent. C'est un son qui a une représentation, une cause et un contexte. Par exemple, dans le contexte des transports, le fait de faire une queue de poisson (la cause) peut amener un énervement d'un conducteur qui sera représenté par un son de klaxon.
- les marqueurs sonores, **soundmarks** qui correspondent à une catégorie intermédiaire des deux précédentes. Il s'agit d'un son qui se réfère à une communauté, qui possède certaines qualités qui le rendent unique, remarquable. Citons par exemple le son des cloches de Big Ben à Londres.

Partant de ces constatations, il me semble intéressant d'enrichir la description automatique de l'environnement sonore.

## Environnement sonore

Bien que l'ensemble des travaux présentés jusqu'ici puissent alimenter la description de l'environnement sonore, il manque indéniablement une notion de **niveau** (ou plan) dans les unités sonores.

### Niveau (ou échelle) de l'environnement sonore

Le niveau n'est pas forcément vu comme une unité du volume (énergie) ou de structuration mais plutôt une indication sur les superpositions des sons et leur importance respective.

Prenons un exemple : *dans la campagne, deux personnes discutent lorsque deux coups de feu retentissent...*

Il serait intéressant de segmenter les événements en différents plans sonores : en arrière plan les sons typiques de l'environnement (dans cet exemple, des oiseaux ou des vaches pourraient correspondre aux *keynote sounds*), en premier plan le son qui capte directement notre attention

(ici, les coups de feu comme *signal sounds*). Dans un plan intermédiaire des sons qui peuvent aussi bien être pris pour un arrière plan que pour un premier plan suivant le contexte (la parole entre les deux personnes peut être considérée comme un premier plan avant que les coups de feu interviennent).

L'intérêt de cette représentation est qu'il n'est pas forcément nécessaire d'identifier chacun des sons présents pour caractériser un environnement sonore mais de segmenter en ces différents niveaux.

Sur le plan scientifique, il convient sans doute de créer une méthode qui permette de fusionner différentes approches :

- une partie non-supervisée pourrait permettre de segmenter (et éventuellement regrouper) les fonds sonores,
- une seconde supervisée devrait identifier des sons connus.

Il est important de signaler que nous ne partons pas de rien ! En effet, des méthodes de séparation de sources, d'extraction de mélodie principale, de segmentation (type PMB ou SRL) existent et peuvent être d'un grand intérêt pour délimiter l'arrière plan sonore. De plus, n'importe quelle méthode de détection audio (telle celles que nous venons de voir dans ce document) peut enrichir le premier plan, que ce soit sur le contenu de parole, de musique ou de bruit.

À travers le projet ANR Corpus CIESS [23], nous sommes en train d'initier ce travail.

Pour l'instant il s'agit, dans une ville (rue de Toulouse semi-piétonnière), de repérer les événements acoustiques qui se détachent du fond sonore en vue de les identifier (tagger) automatiquement si possible (le label pourrait être ajouté de manière manuelle dans un second temps). Ceci doit nous permettre de caractériser l'environnement urbain et de pouvoir le modifier au besoin (exemple : simuler la transformation d'une rue avec voitures en une rue piétonnière).

Cette manière de fonctionner devrait nous amener plus de cohérence et de complémentarité entre le traitement et la synthèse des sons.

Compte tenu de mon expérience et surtout de celle de mon équipe, nous pensons nous attaquer au problème sous l'angle de la segmentation par des approches statistiques : l'algorithme de DFB [André-Obrecht 88] et le critère BIC devraient être encore explorés afin d'arriver à des unités homogènes.

Néanmoins, concernant les différents niveaux permettant de décrire l'environnement sonore (en quelque sorte une **échelle de l'environnement sonore**), quelques paramètres me semblent d'une grande pertinence.

Citons tout d'abord *l'énergie du signal*. Il paraît assez logique que celle-ci soit fortement corrélée à l'échelle sonore : un arrière-plan avec une énergie trop importante devrait être perçu comme un premier plan. Ce paramètre est d'autant plus intéressant qu'il est facile à estimer. Le point délicat réside quand même à délimiter la zone correspondante.

23. http://petra.univ-tlse2.fr/ciess/

Dans le même ordre d'idée, *la durée de l'événement sonore* peut donner une indication sur le plan auquel il appartient. En effet, le fond sonore est, en général, d'une durée assez longue alors que les éléments de premier plan sont plutôt brefs.

Ensuite, il me semble que *l'intelligibilité* du signal est un autre point clé. Quelque chose qui ne sera pas perçu distinctement doit avoir un niveau beaucoup plus en retrait que quelque chose de clair. Bien qu'elles soient assez limitées, des mesures telles le RASTI (*RApid Speech Transmission Index*), pourraient évaluer cette intelligibilité.

Pour aller plus loin, il pourrait être intéressant de proposer une *mesure de la compréhension*. Comme pour l'intelligibilité, la compréhension devrait donner des indications sur le niveau de l'événement sonore analysé.

Nous sommes d'ailleurs en train de mettre en place une telle mesure pour la compréhension de la parole dans un projet région Midi-Pyrénées AGILE IT. Ainsi, à travers des systèmes de Décodage Acoustico-Phonétique (DAP), de Transcription et de Compréhension, nous cherchons à évaluer comment le message est assimilé. Afin de proposer une échelle de compréhension, nous construisons un système global qui puissent rendre compte des scores obtenus par un humain dans un contexte de perception [Fontan 12].

Ce type de méthode pourrait permettre, dans le cas de la parole, de distinguer un effet « babble » (cocktail party) en arrière plan d'une conversation.

L'idéal serait que cette mesure soit généralisée aux autres composantes primaires (musique et bruit). Bien évidemment, ceci amène d'énormes difficultés. D'une part, les différents systèmes mis en place en parole, n'existent pas forcément pour les autres catégories. D'autre part, la superposition des unités sonores (parole, musique, bruit) va poser des problèmes d'identification...

Ce qui vient d'être décrit semble néanmoins très ambitieux pour être mis en place rapidement.

### Un peu plus de pragmatisme

À moyen terme et comme déjà indiqué dans la discussion de la segmentation sonore, il semble intéressant de détecter, à chaque instant, le nombre de sources dans un environnement sonore.

Actuellement deux sujets axés sur la musique sont déjà amorcés.

Le premier fait suite aux travaux de Maxime Le Coz sur la détection des superpositions de sources, vue en section 3.1.1 : il apparaît, en effet, raisonnable de **détecter le tuilage** aussi bien en parole, qu'en musique. Le tuilage peut être défini par une zone de superposition de deux entités $E_1$ et $E_2$ (aussi bien des locuteurs, que des chanteurs ou des instruments), sachant qu'avant la superposition, seule l'entité $E_1$ est présente et qu'après, seule l'entité $E_2$ reste. Les entités sont singulières (un locuteur ou un chanteur) ou plurielles (un groupe de locuteurs ou un groupe de chanteurs).

La figure 1 est une illustration de ce phénomène avec deux locuteurs.

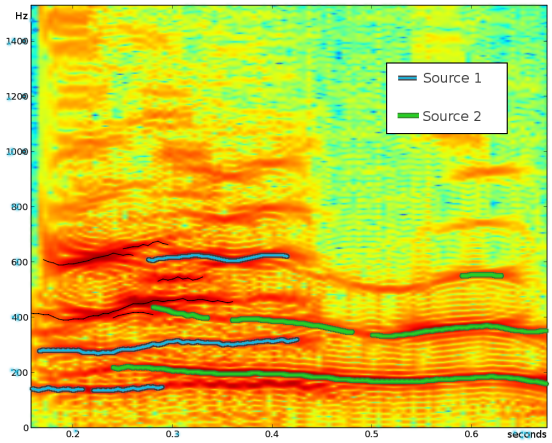



FIGURE 1 – Exemple du phénomène de tuilage en parole. Le suivi des fréquences avant et après la zone de superposition permet de mettre en évidence l'existence des deux entités.

Notre idée est dans un premier temps de détecter la zone de superposition, puis de montrer que dans cette zone, il existe une source qui est présente avant la superposition et une autre source qui est présente après. Une analyse temporelle de la continuité des suivis fréquentiels sera alors effectuée.

Le second sujet est « dans l'air du temps » car divers laboratoires s'y intéressent sans que pour l'instant des résultats concrets émergent. Il concerne la **segmentation en tours de chant**. Par analogie à la segmentation en tours de parole, la segmentation en tours de chant consiste à déterminer les instants de changement de chanteurs. Il s'agit d'un problème indépendant de la détection chant/non-chant vue dans la section 2.2, car un même chanteur peut naturellement faire une pause pendant son chant, alors que l'enchaînement des chanteurs se fait souvent sans pause, voire avec superposition des voix et des instruments.

Dans la lignée de la segmentation en locuteurs de la section 1.1, nous avons débuté cette étude par une implémentation du critère BIC. L'exploration porte à l'heure actuelle sur deux éléments :

- la valeur du coefficient de pénalité $\lambda$ est revisitée du fait de la taille des segments : en général, plus longs,
- la taille de la fenêtre sur laquelle nous appliquons le critère BIC sera nécessairement plus grande. Là aussi des segments plus longs doivent être favorisés.

Les premiers résultats, dans le contexte de chant *a cappella* sur des enregistrements difficiles du corpus DIADEMS[24], sont encourageants. Bien évidemment, quand des instruments sont présents, de nombreuses sur-segmentations apparaissent. Il semble nécessaire de faire évoluer les descripteurs acoustiques afin qu'ils représentent plus finement la musique. D'autres métriques de détection devront sans doute aussi être abordées. Maroua Thlithi débute une thèse que je co-encadre avec Régine André-Obrecht sur ce sujet.

24. http://diadems.telemeta.org/

## Contexte sonore et visuel

Même si la majorité de mes travaux actuels touchent davantage la composante sonore que visuelle, je reste néanmoins intéressé par le couplage de l'audio et de la vidéo.

D'une part, de nombreuses évolutions sont directement possibles sur les travaux que j'ai présentés dans ce manuscrit : quelques pistes ont d'ailleurs été lancées dans la discussion sur la segmentation audiovisuelle en section 5.3.2.

D'autre part, j'ai été touché par ma collaboration avec le milieu médical dans le projet IMMED sur la segmentation en activités (voir section 4.4), peut-être, parce que cela me donne l'impression d'effectuer une recherche plus « vitale »...

### Coup de cœur

Effectivement, j'essaye avec mes collègues bordelais (Jenny Benois-Pineau et Rémi Mégret) de continuer notre collaboration dans le domaine médical.

Notre ambition est de travailler sur l'analyse du contexte des chutes de patients atteints de la maladie de Parkinson. Le nerf de la guerre étant souvent l'argent, nous sommes en train de déposer des projets à différents niveaux (local, régional, national) afin d'obtenir un financement et ainsi débuter cette étude.

Le contexte est le suivant : à ce jour, il n'existe pas de thérapeutique capable de prévenir les chutes et leur prévention ne peut que s'appuyer sur des stratégies comportementales, de rééducation et de protection simple. Notre solution est alors de proposer, via un dispositif portable, des capteurs de mouvement pour la détection des chutes et l'analyse du mouvement permettant de reconnaître les signes de fluctuations motrices (tremblements, bradykinésie, dyskinésie), et des capteurs audio/vidéo portés permettant de capturer le contexte de ces chutes et d'y détecter les causes de celles-ci.

L'analyse du contexte des chutes pourrait être appréhendée selon trois axes :

- la caractérisation du mouvement précédant la chute et la détection de signes avant-coureurs spécifiques tels que la bradykinésie,
- la caractérisation de l'environnement et des activités menées par le patient,
- la détection d'événements sonores pouvant expliquer une perte d'attention expliquant la chute.

Bien évidemment, mes contributions se situent sur la dernière partie ainsi que sur la fusion des trois axes qui permettent d'obtenir le contexte audio-vidéo. L'expérience acquise avec les HMM y sera sans nul doute très utile.

### Généralisation de l'environnement

Si nous arrivons à caractériser l'environnement sonore, pourquoi ne pas aller plus loin en l'élargissant à la vision et ainsi caractériser l'environnement audiovisuel ?

La prise en compte du contexte (environnement) est en effet un grand manque dans les méthodes actuelles. En effet, soit la majorité des méthodes le simplifie en construisant un système par contexte, soit elle s'en abstient en utilisant des normalisations.

Ce travail a une visée à très long terme. Néanmoins, une première étape pourrait être de **détecter des scènes (ou objets) audiovisuelles**. Ceci pourra s'effectuer en trois temps : segmentation en sources sonores et visuelles, regroupement mono-média et regroupement multimédia. Les deux premières étapes sont actuellement assez matures.

En reprenant notre idée de IN, OUT, OFF sur les intervenants (cf. section 4.2), nous pourrions la généraliser à tout objet, dans l'optique de fusionner les éléments sonores et visuels.

Prenons un exemple pour se convaincre de l'intérêt. À partir d'un enregistrement vidéo situé place du Capitole (voir figure 2), l'idée serait de pouvoir :

- attribuer à chaque vélo de l'image et à chaque véhicule motorisé, le son qui lui correspond,
- attribuer à chaque personne, sa voix éventuelle,
- détecter des objets sonores qui n'ont pas de correspondance visuelle,
- détecter des objets visuels n'ont pas de correspondance sonore,
- etc.

FIGURE 2 – Image représentant l'intérêt d'une détection d'objets sonores et visuels en vue d'une caractérisation de l'environnement audiovisuel.

# Bibliographie


[Aha 95] D.W. Aha & R. L. Bankert. *A comparative evaluation of sequential feature selection algorithms.* In Proceedings of the 5th International Workshop on Artificial Intelligence and Statistics, pages 199–206, 1995.

[Albiol 04] A. Albiol, L. Torrest & E. J. Delp. *Two are better than one : when audio comes to the rescue of video.* In in Proceedings of the 5th European Workshop on Image Analysis for Multimedia Interactive Services, Lisbonne, Portugal, April 2004.

[André-Obrecht 93] R. André-Obrecht. *Segmentation et parole ?* Thèse d'état, IRISA, 1993.

[André-Obrecht 88] R. André-Obrecht. *A New Statistical Approach for Automatic Speech Segmentation.* IEEE Transactions on Audio, Speech, and Signal Processing, vol. 36, no. 1, pages 29–40, January 1988.

[Arias 04] J. A. Arias. Méthodes à vecteurs de support et indexation sonore. Rapport de DEA, IRIT, Université Paul Sabatier, Toulouse III, June 2004.

[Arroabarren 07] I. Arroabarren & A. Carlosena. *Voice Production Mechanisms of Vocal Vibrato in Male Singers.* IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 1, pages 320–332, January 2007.

[Atal 76] B.S. Atal & L. Rabiner. *A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition.* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 24, no. 3, pages 201–212, 1976.

[Bar-Hillel 05] A. Bar-Hillel, T. Hertz, N. Shental & D. Weinshall. *Learning a Mahalanobis Metric from Equivalence Constraints.* J. Mach. Learn. Res., vol. 6, pages 937–965, December 2005.

[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.

[Bazillon 11] T Bazillon, B. Maza, M. Rouvier, F. Bechet & A. Nasr. *Speaker Role Recognition using question detection and characterization.* In Proceedings of Interspeech, pages 1333–1336, 2011.

[Bigot 11] B. Bigot. *Recherche du rôle des intervenants et de leurs interactions pour la structuration de documents audiovisuels.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France, July 2011.

[Bragg 20] Sir W. H. Bragg. The World of Sound. G. Bell and Sons ltd., London, 1920.

[Brassard 96] G. Brassard & P. Bratley. Fundamentals of algorithmics. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[Canseco-Rodriguez 04] L. Canseco-Rodriguez, L. Lamel & J-L. Gauvain. *Towards using STT for Broadcast News Speaker Diarization.* In Proceedings of DARPA RT04, 2004.

[Chen 98] S. S. Chen & P. S. Gopalakrishnan. *Speaker, Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion.* In DARPA Speech Recognition Workshop, pages 127–132, 1998.

[Chen 05] J. Chen, A. H. Kam, J. Zhang, N. Liu & L. Shue. *Bathroom activity monitoring based on sound.* In Proceedings of the Third international conference on Pervasive Computing, PERVASIVE'05, pages 47–61. Springer-Verlag, 2005.

[Comon 94] P. Comon. *Independent component analysis, a new concept ?* Signal Processing, vol. 36, no. 3, pages 287–314, April 1994.

[Cortes 95] C. Cortes & V. Vapnik. *Support-Vector Networks.* Machine Learning, vol. 20, no. 3, pages 273–297, 1995.

[Cox 94] T. Cox & M. Cox. Multidimensional scaling. Chapman & Hall, London, 1994.

[de Cheveigné 02] A. de Cheveigné & H. Kawahara. *YIN, a Fundamental Frequency Estimator for Speech and Music.* Journal of the Acoustical Society of America, vol. 111, no. 4, pages 1917–1930, 2002.

[Dempster 77] A. P. Dempster, N. M. Laird & D. B. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, vol. 39 (Series B), pages 1–38, 1977.

[Devijver 82] P. A. Devijver & J. Kittler. Pattern recognition : A statistical approach. Prentice Hall, 1982.

[Dovgalecs 11] V. Dovgalecs. *Indoor location estimation using a wearable camera with application to the monitoring of persons at home.* PhD thesis, IMS, Université Sciences et Technologies, Bordeaux I, December 2011.

[Duda 01] R. O. Duda, P. E. Hart & D. G. Stork. Pattern Classification (2nd Edition). Wiley-Interscience Publication, 2001.

[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.

[Dufour 11] R. Dufour, Y. Estève & P. Deléglise. *Investigation of Spontaneous Speech Characterization Applied to Speaker Role Recognition.* In Proceedings of Interspeech, pages 917–920, 2011.

[El Khoury 06] E. El Khoury. Segmentation et regroupement en locuteurs d'un document sonore. Rapport de master, Toulouse, France, June 2006.

[El Khoury 07] E. El Khoury, C. Senac & R. André-Obrecht. *Speaker Diarization : Towards a more Robust and Portable System.* In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 489–492. IEEE, 2007.

[El Khoury 09] E. El Khoury, C. Senac & J. Pinquier. *Improved Speaker Diarization System for Meetings.* In International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 4241–4244. IEEE, 2009.

[El Khoury 10] E. El Khoury. *Unsupervised Video Indexing based on Audiovisual Characterization of Persons.* Thèse de doctorat, Université de Toulouse, Toulouse, France, June 2010.

[Estève 10] Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet & J. Farinas. *The EPAC Corpus : Manual and Automatic Annotations of Conversational Speech in French Broadcast News.* In LREC. European Language Resources Association, 2010.

[Fanelli 13] G. Fanelli, M. Dantone, J. Gall, A. Fossati & L. Van Gool. *Random Forests for Real Time 3D Face Analysis.* International Journal of Computer Vision, vol. 101, no. 3, pages 437–458, August 2013.

[Fisher 36] R. A. Fisher. *The Use of Multiple Measurements in Taxonomic Problems.* Annals of Eugenics, vol. 7, no. 7, pages 179–188, 1936.

[Fogarty 06] James Fogarty. *Sensing from the basement : a feasibility study of unobtrusive and low-cost home activity recognition.* In In Proceedings of the 19th Annual ACM Symposium on User interface Software and Technology UIST 2006, pages 91–100, 2006.

[Fontan 12] L. Fontan. *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication.* PhD thesis, Université Toulouse le Mirail - Toulouse II, November 2012.

[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 INTERSPEECH, Conference of the International Speech Communication Association, pages 1149–1152. ISCA, 2005.

[Galliano 09] S. Galliano, G. Gravier & L. Chaubard. *The ester 2 evaluation campaign for the rich transcription of French radio broadcasts.* In INTERSPEECH, Conference of the International Speech Communication Association, pages 2583–2586. ISCA, 2009.

[Gaver 93] W. W. Gaver, R. L. Jenison, C. L. Schmidt & J. G. Neuhoff. *How do we hear in the world ? Explorations in ecological acoustics* . Ecological-Psychology, vol. 5, no. 4, pages 285–313, 1993.

[Geffen 11] M. N. Geffen, J. Gervain, J. F. Werker & M. O. Magnasco. *Auditory perception of self-similarity in water sounds.* Frontiers in Integrative Neuroscience, vol. 5, page 15, 2011.

[Geng 05] X. Geng, D. Zhan & Z. Zhou. *Supervised nonlinear dimensionality reduction for visualization and classification.* IEEE Transactions on systems, man, and cybernetics-part B : cybernetics, vol. 35, pages 1098–1107, 2005.

[Gerhard 02] David B. Gerhard. *Perceptual Features for a Fuzzy Speech-Song Classification.* In International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 4, pages 4160–4163. IEEE, May 2002.

[Gianni 07] F. Gianni, J. Pinquier & E. Kijak. *ACADI Showcase - Automatic Character Indexing in Audiovisual Document.* In ACM International Conference on Image and Video Retrieval (CIVR). ACM, 2007.

[Gish 91] H. Gish, M. H. Siu & R. Rohlicek. *Segregation of speakers for speech recognition and speaker identification.* In International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 873–876. IEEE, May 1991.

[Guyon 03] I. Guyon. *An introduction to variable and feature selection.* Journal of Machine Learning Research, vol. 3, pages 1157–1182, 2003.

[Guyot 12] P. Guyot, J. Pinquier & R. André-Obrecht. *Water flow detection from a wearable device with a new feature, the spectral cover (regular paper).* In International Workshop on Content-Based Multimedia Indexing (CBMI), pages 139–142. IEEE, June 2012.

[Guyot 13] P. Guyot, J. Pinquier & R. André-Obrecht. *Water sound recognition based on physical models.* In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), June 2013.

[Gygi 07] B. Gygi, G.R. Kidd & C. S. Watson. *Similarity and categorization of environmental sounds.* Perception And psychophysics, vol. 69, no. 6, pages 839–55, 2007.

[Haeb-Umbach 92] R. Haeb-Umbach & H. Ney. *Linear discriminant analysis for improved large vocabulary continuous speech recognition.* In Proceedings of the IEEE international conference on Acoustics, speech and signal processing, ICASSP'92, pages 13–16. IEEE Computer Society, 1992.

[Haidar 05] S. Haidar. *Comparaison des Documents Audiovisuels par Matrice de Similarité.* PhD thesis, Université Paul Sabatier, Toulouse III, September 2005.

[Heittola 11] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen & Antti Eronen. *Sound event detection in multisource environments using source separa-*

*tion*. In Workshop on machine listening in Multisource Environments, pages 36–40, 2011.

[Hoi 06] S. C. H. Hoi, W. Liu, M. R. Lyu & W. Ma. *Learning distance metrics with contextual constraints for image retrieval*. In Proc. Computer Vision and Pattern Recognition, pages 2072–2078. Murray Hill, 2006.

[Houtgast 85] T. Houtgast & J. M. Steeneken. *A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria*. Journal of the Acoustical Society of America, vol. 77, no. 3, pages 1069–1077, 1985.

[Jaffré 04] G. Jaffré, P. Joly & S. Haidar. *The SAMOVA Shot Boundary Detection for TRECVID Evaluation 2004*. In TRECVID 2004 Workshop, pages 179–183, Gaithersburg, Maryland USA, November 2004. NIST.

[Jaffré 05] Gaël Jaffré. *Indexation de la vidéo par le costume* . Thèse de doctorat, Université Paul Sabatier, Toulouse, France, November 2005.

[Johnston 98] J. D. Johnston. *Transform coding of audio signals using perceptual noise criteria*. IEEE Journal on Selected Areas in Communications, vol. 6, no. 2, pages 314–323, February 1998.

[Karaman 10] S. Karaman, J. Benois-Pineau, R. Megret, V. Dovgalecs, J.-F. Dartigues & Y. Gaestel. *Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases*. In International Conference on Pattern Recognition, pages 4113–4116, Washington, DC, USA, 2010. IEEE Computer Society.

[Karaman 11] S. Karaman. *Indexing of Activities in Wearable Videos : Application to Epidemiological Studies of Aged Dementia*. PhD thesis, LaBRI, Université Sciences et Technologies, Bordeaux I, December 2011.

[Karaman 12] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Megret, J. Pinquier, R. André-Obrecht, Y. Gaestel & J.-F. Dartigues. *Hierarchical Hidden Markov Model in Detecting Activities of Daily Living in Wearable Videos for Studies of Dementia*. Multimedia Tools and Applications, pages 1–29, 2012.

[Kelley 99] C. T. Kelley. *Iterative Methods for Optimization*. Frontiers in Applied Mathematics, vol. SIAM, 1999.

[Klapuri 06] A. Klapuri. *Multiple fundamental frequency estimation by summing harmonic amplitudes*. In International Society for Music Information Retrieval (ISMIR), pages 216–221, 2006.

[Kraaij 04] W. Kraaij, A. F. Smeaton, P. Over & J. Arlandis. *TRECVID 2004 – An introduction*. In TREC Video Retrieval Evaluation Online Proceedings, 2004.

[Lachambre 07] H. Lachambre, R. André-Obrecht & J. Pinquier. *Singing Voice Characterization for Audio Indexing*. In 15th European Signal Processing Conference (EUSIPCO), pages 1563–1540, 2007.

[Lachambre 09] H. Lachambre. *Caractérisation de l'environnement musical dans les documents audiovisuels.* Thèse de doctorat, Université de Toulouse, Toulouse, France, December 2009.

[Lachambre 11] H. Lachambre, J. Pinquier & R. André-Obrecht. *Distinguishing Monophonies from Polyphonies using Weibull Bivariate Distributions.* IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 6, pages 1837–1842, août 2011.

[Le Coz 10] M. Le Coz, H. Lachambre, L. Koenig & R. André-Obrecht. *A Segmentation-based Tempo Induction Method (regular paper).* In International Society for Music Information Retrieval Conference (ISMIR), pages 27–31. Dagstuhl Research Online Publication Server, August 2010.

[Le Coz 12] M. Le Coz, R. André-Obrecht & J. Pinquier. *Feasibility of the Detection of Choirs for Ethnomusicologic Music Indexing.* In International Workshop on Content-Based Multimedia Indexing (CBMI), pages 145–148. IEEE, June 2012.

[Le Coz 13] M. Le Coz, J. Pinquier, R. André-Obrecht & J. Mauclair. *Audio Indexing Including Frequency Tracking of Simultaneous Multiple Sources in Speech and Music.* In International Workshop on Content-Based Multimedia Indexing (CBMI), pages 23–25. IEEE, juin 2013.

[Leighton 97] T.G. Leighton. The acoustic bubble. Academic Press, 1997.

[Li 01] D. Li, G. Wei, I. K. Sethi & N. Dimitrova. *Fusion of visual and audio features for person identification in real video*, 2001.

[Liu 06] Y. Liu. *Initial study on automatic identification of speaker role in broadcast news speech.* In Proeedings of Human Language Technology Conference of the NAACL, pages 81–84, 2006.

[Lu 90] J.C. Lu & G.K. Bhattacharyya. *Some New Constructions of Bivariate Weibull Models.* Annals of Institute of Statistical Mathematics, vol. 42, no. 3, pages 543–559, 1990.

[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.

[Mauclair 03] J. Mauclair. Fusion de paramètres pour une classification Parole/Musique/Bruit. Rapport de DEA, IRIT, Université Paul Sabatier, Toulouse III, June 2003.

[Mcheik 06] A. Mcheik. Application des matrices de similarité à la comparaison de contenus sonores. Rapport de master, Toulouse, France, June 2006.

[Meron 00] Y. Meron & K. Hirose. *Synthesis of Vibrato Singing.* In International Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 2, pages 745–748, 2000.

[Mégret 10] R. Mégret, V. Dovgalecs, H. Wannous, S. Karaman, J. Benois-Pineau, E. El Khoury, J. Pinquier, P. Joly, R. André-Obrecht, Y. Gaëstel & J.-F. Dartigues. *The IMMED project : wearable video monitoring of people with age dementia.* In Proceedings of the international conference on Multimedia, MM'10, pages 1299–1302, New York, NY, USA, 2010. ACM.

[Minnaert 33] M. Minnaert. *On musical air-bubbles and the sound of running water.* Philosophical Magazine, vol. 16, no. 103, pages 235–248, 1933.

[Moddemeijer 89] R. Moddemeijer. *On Estimation of Entropy and Mutual Information of Continuous Distributions.* Signal Processing, vol. 16, no. 3, pages 233–246, 1989.

[Narendra 77] P. M. Narendra & K. Fukunaga. *A Branch and Bound Algorithm for Feature Subset Selection.* IEEE Transactions Computing, vol. 26, no. 9, pages 917–922, September 1977.

[Ohish 05] Y.i Ohish, M. Goto, K. Itou & K. Takeda. *Discrimination between singing and speaking voices.* In Interspeech - European Conference on Speech Communication and Technology, pages 1141–1144, September 2005.

[Pachet 07] F. Pachet & P. Roy. *Exploring Billions of Audio Features.* In Content-Based Multimedia Indexing (CBMI), pages 227–235, 2007.

[Pearson 01] K. Pearson. *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, vol. 2, no. 6, pages 559–572, 1901.

[Peeters 03] G. Peeters. *Automatic Classification of Large Musical Instrument Databases Using Hierachical Classifiers with Inertia Ratio Maximization.* In 115th AES Convention, 2003.

[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.

[Philippeau 05] J. Philippeau. Caractérisation d'un intervenant dans un document audiovisuel. Rapport de stage, Toulouse, France, June 2005.

[Philippeau 09] J. Philippeau. *Apprentissage de similarités pour l'aide à l'organisation de contenus audiovisuels.* Thèse de doctorat, Université de Toulouse, Toulouse, France, June 2009.

[Pinquier 03a] J. Pinquier, J.-L. Rouas & R. André-Obrecht. *Fusion de paramètres pour une classification automatique parole/musique robuste.* Technique et Science Informatiques (TSI), vol. 22, no. 7-8, pages 831–852, 2003.

[Pinquier 03b] J. Pinquier, Jean-Luc Rouas & R. André-Obrecht. *A Fusion Study in Speech / Music Classification.* In IEEE International Conference on Audio, Speech and Signal Processing, Hong-Kong, China, April 2003.

[Pinquier 04] J. Pinquier. *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle.* Thèse de doctorat, Université Paul Sabatier, Toulouse, France, December 2004.

[Pérès 08] K. Pérès, C. Helmer, H. Amieva, J.-M. Orgogozo, I. Rouch, J.-F. Dartigues & P. Barberger-Gateau. *Natural history of decline in instrumental activities of daily living performance over the 10 years preceding the clinical diagnosis of dementia : a prospective population-based study.* Journal of the American Geriatrics Society, vol. 56, no. 1, pages 37–44, 2008.

[Rabiner 89] L. R. Rabiner, J. G. Wilpon & F. K. Soong. *High Performance Connected Digit Recognition using Hidden Markov Models.* IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 8, pages 1214–1225, 1989.

[Schafer 77] R.M. Schafer. The tuning of the world. Borzoi book. Knopf, 1977.

[Seashore 38] C. E. Seashore. Psychology of Music. McGraw-Hill Book Company, inc., 1938.

[Signol 09] F. Signol. *Estimation de fréquences fondamentales multiples en vue de la séparation de signaux de parole mélangés dans un même canal.* PhD thesis, Université Paris Sud - Paris XI, December 2009.

[Taati 10] B. Taati, J. Snoek, D. Giesbrecht & A. Mihailidis. *Water Flow Detection in a Handwashing Task.* In Conference on Computer and Robot Vision (CRV), pages 175–182, 2010.

[Taniguchi 05] Toru Taniguchi, Akishige Adachi, Shigeki Okawa, Masaaki Honda & Katsuhiko Shirai. *Discrimination of Speech, Musical Istruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals.* In Interspeech - European Conference on Speech Communication and Technology. ISCA, September 2005.

[Tenenbaum 00] J. B. Tenenbaum, V. de Silva & J. C. Langford. *A Global Geometric Framework for Nonlinear Dimensionality Reduction.* Science, vol. 290, no. 5500, pages 2319–2323, December 2000.

[Timmers 00] R. Timmers & P. Desain. *Vibrato : Questions and Answers from Musicians and Science.* In Proc. Int. Conf. on Music Perception and Cognition, 2000.

[Tsai 08] W.-H. Tsai, S.-J. Liao & C. Lai. *Automatic Identification of Simultaneous Singer Recordings.* In Proc. of the 9th International Conference on Music Information Retrieval (ISMIR'08), pages 115–120, 2008.

[Tsekeridou 01] S. Tsekeridou & I. Pitas. *Content-based video parsing and indexing based on audio-visual interaction.* Circuits and Systems for Video Technology, IEEE Transactions on, vol. 11, no. 4, pages 522–535, 2001.

[van Dam 97] A. van Dam. *Post-WIMP User Interfaces*, 1997.

[Vapnik 99] V. Vapnik. The Nature of Statistical Learning Theory. Springer, 1999.

[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.

[Viola 01] P. Viola & M. Jones. *Rapid object detection using a boosted cascade of simple features*. In Computer Vision and Pattern Recognition (CVPR), volume 1, pages 511–518, 2001.

[Young 94] S. Young. *The HTK Hidden Markov Model Toolkit : Design and Philosophy*. Rapport technique 152, Cambridge University Engineering Department, UK, 1994.

[Zhang 97] H. J. Zhang, C. Y. Low, S. W. Smoliar & J. H. Wu. *Video Parsing, Retrieval and Browsing - An Integrated and Content-Based Solution*. In Intelligent multimedia information retrieval, pages 139–158, 1997.