Indexation cross-média vidéo/son des contenus multimédia numériques

Résumé : Dans ce travail de thèse, nous nous intéressons au problème de l'indexation des flux audio-visuels numériques en scènes ou "chapitres sémantiques" à partir des informations extraites des flux audio et vidéo.

Les travaux menés dans ce domaine de recherche sont récents, notamment les approches basées sur l'analyse cross-média (audio et vidéo). Dans cette thèse, nous proposons une méthode générique de détection des frontières de scènes dans les contenus télédiffusés basée sur une approche de décision statistique.

Puis, après avoir étudié les transitions audio qui caractérisent ces frontières de scènes, nous proposons une approche pour la caractérisation des transitions entre les bruits. Nous étudions divers descripteurs du signal audio pour sélectionner les plus pertinents d'entre eux pour la caractérisation des sons bruités.

Enfin, une méthode statistique de classification des sons bruités dans le but d'affiner la méthode de segmentation précédente est proposée. Elle permet de caractériser les transitions entre les bruits au sein d'une même classe en sélectionnant les descripteurs appropriés.

Discipline : Informatique

Mots-clés : indexation cross-média, décision bayésienne, scènes sémantiques, analyse de sons bruités, segmentation et classification de sons bruités.

Cross-media approach for semantic partitioning of digital multimedia content

Abstract: In this work, the problem of indexing a digital audio-visual content into scenes or "semantic chapters" by taking into consideration the information extracted from both audio and video streams is considered.

There has been considerable prior recent work in this area, particularly in the areas dealing with cross-media (audio and video). In this work, a generic method to characterize scene borders in broadcasting audio-visual contents is proposed. It is based on a statistical decision model.

More precisely, audio transitions related to a scene border are considered. An approach to characterize noise-to-noise transitions is developed. Various low-level audio descriptors are studied in order to select the ones which best characterize noisy sounds.

Finally, a statistical approach to classify noisy sounds is developed in order to enhance previous segmentation methods. It allows, by selecting appropriate descriptors, to detect the noise-to-noise transitions occurring inside the same noise class.

Discipline: Computer-Science

Keywords: cross-media indexing, Bayesian decision model, semantics scenes, noisy sounds analysis, noisy sounds segmentation and classification.

LaBRI,
Université Bordeaux 1,
351 cours de la Libération,
33405 Talence Cedex (FRANCE).

N° d'ordre : 3136

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX I

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par **Nicolas LOUIS**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Indexation cross-média vidéo/son des contenus multimédia numériques

---

Soutenue le : 06 Février 2006

Après avis des rapporteurs :

Régine André-Obrecht	Professeur
François Pachet	Chercheur chez Sony CSL

Devant la commission d'examen composée de :

Henri Nicolas	Professeur	Président
Régine André-Obrecht	Professeur	Rapporteur
François Pachet	Chercheur chez Sony CSL	Rapporteur
Myriam Desainte-Catherine	Professeur	Examinateur
Jenny Benois-Pineau	Professeur	Examinateur
Jan Nesvadba	Chercheur chez Philips Research	Examinateur

– 2006 –

Remerciements

... au hasard de la vie ...

Mener à bien un travail de thèse n'est pas chose facile, bien des personnes l'ayant vécu peuvent en témoigner. Pour réussir une thèse il faut faire preuve à la fois de volonté, d'opiniâtreté et avoir cette "fameuse soif" de curiosité.

Alors que je terminais ma maîtrise d'informatique, j'obtenais tout juste l'autorisation de suivre un DEA d'informatique. Au passage j'en profite pour remercier chaleureusement M. Serge Dulucq, actuel directeur du Laboratoire Bordelais de Recherche en Informatique (LaBRI), pour la confiance qu'il m'a accordée en m'acceptant en DEA. Tout commença lorsque, dans le cadre d'un exercice de bibliographie, je devais choisir un article pour le présenter. Mon choix s'est alors porté sur un article proposé par un membre du laboratoire que je ne connaissais pas et pour cause, il venait tout juste d'intégrer l'équipe Image et Son du LaBRI en tant que Professeur. Lorsque je fis la connaissance de cette personne, j'étais loin de m'imaginer qu'allait naître une collaboration aussi enrichissante.
C'est lorsqu'elle me présenta son sujet de DEA que je lui annonçais que j'avais déjà contacté un autre enseignant pour un sujet portant sur le domaine du son. Qu'à cela ne tienne, je reçus le lendemain même un courriel m'indiquant qu'un nouveau sujet avait été rédigé conjointement par ces deux enseignants. Ces deux personnes qui allaient devenir mes directrices de thèse pendant plus de 3 ans sont Pr. Jenny Benois-Pineau et Pr. Myriam Desainte-Catherine, deux femmes de grand talent faisant de leur métier une passion qu'elles s'efforcent de transmettre à des générations d'étudiants d'année en année.
Le sujet portait sur l'analyse des vidéos numériques par des méthodes combinant le traitement des flux audio et vidéo.
Lorsque je suis arrivé en DEA mes connaissances était limitées dans les domaines liés à mon sujet de thèse. Néanmoins, je n'oublierai jamais les nombreuses réunions fructueuses que nous avons eu avec mes directrices de thèse qui ont toujours su se rendre disponibles pour combler mes lacunes.

Mon DEA en poche, il me fallait obtenir une bourse pour poursuivre une thèse. Une fois de plus, je ne remercierai jamais assez Jenny et Myriam pour les efforts consentis auprès de la direction du DEA et pour avoir monté un dossier auprès de la région Aquitaine afin obtenir finalement ma bourse pour une durée de 3 ans.
Puis, tout s'est enchaîné si vite, une visite de Jan Nesvadba travaillant chez Philips Research et me voilà parti 10 mois, dans le cadre de ma thèse, à Eindhoven (Pays-Bas) encadré par Jan qui m'a beaucoup apporté tant sur la plan humain que professionnel. C'est avec plaisir que Jan fait partie de mon jury de thèse et je profite de cette occasion pour le remercier infiniment pour tout ce qu'il m'a apporté. (Jan let me tell you how much you help me to improve myself.

I would like to thank you so much for all you have done for me and for the chance you gave me with Jenny and Myriam to acquire such experience at Philips Research.)
Dès mon retour à Bordeaux, j'ai eu l'occasion d'aller à Barcelone pour assister à la conférence ISMIR 2004 au cours de laquelle j'ai rencontré M. François Pachet Maître de conférence habilité chez Sony-CSL Paris. J'étais, une fois encore, loin de m'imaginer que M. François Pachet allait lui aussi faire partie de mon jury de thèse. Je voudrais donc le remercier à double titre, d'une part pour avoir accepté d'être rapporteur de ma thèse et d'autre part pour avoir lu le manuscrit et rédigé son rapport final aussi rapidement compte tenu de ses multiples activités.
Pour les mêmes raisons, je remercie avec le même enthousiasme Pr. Régine André-Obrecht.
Dans un intervalle de temps réduit et en période de fêtes de fin d'année, mes rapporteurs ont pris le temps de lire mon mémoire et d'y apporter les remarques qui ont permis d'en améliorer la qualité au final.
Enfin je remercie Pr. Henri Nicolas, récemment nommé au LaBRI en tant que Professeur dans l'équipe Image et Son, d'avoir accepté de présider mon jury de thèse.

Je tenais à remercier tout particulièrement Pierre Hanna, maintenant Maître de conférence au LaBRI et brillant capitaine de notre équipe de foot, pour les travaux de recherche que nous avons menés ensemble ainsi que pour son soutien dans les moments difficiles et aussi parce qu'il a toujours répondu présent pour me conseiller au cours de cette thèse. Sans Pierre, ma thèse n'aurait pas été aussi étoffée.
Je n'oublie pas dans ces remerciements mes compagnons cosmopolites de chez Philips : Marzia (Italie), Robert (Hollande), Pedro (Portugal), Jin Yan (Chine) et René (Espagne). Sans oublier mes collègues et néanmoins amis du bureau 219 puis 273 au LaBRI : Laurent à qui je souhaite bon courage pour son futur professionnel, Lionel qui est sur le point de terminer sa thèse, Francesca que je félicite pour sa thèse et Pétra à qui je souhaite bonne réussite pour la suite de sa thèse.
Je tiens aussi à remercier mes amis : Alexis, Vincent, Marie-Vincente, Geoffrey, Thierry, Christine, Nelly, Hugues, Marikel, l'ensemble des Blasteurs, les joueurs de l'équipe de foot de l'Afodib, les joueurs de l'équipe de foot du laboratoire SIC à Poitiers. Tous m'ont encouragé et m'ont permis de m'évader quelque peu lors de ces trois années de thèse.
Je tenais aussi à faire un clin d'œil à Christine Fernandez et Noël Richard du laboratoire SIC de Poitiers pour leurs encouragements et pour m'avoir permis de terminer ma thèse dans les meilleures conditions au sein de SIC.
Une pensée aussi envers le personnel administratif et technique du LaBRI pour leur dévouement et leur aide : Philippe Biais, Jean-Louis Lassartesses, Bernard Duflo, Cathy Roubineau et les membres de l'équipe système.
Avant de terminer ces remerciements, je voudrais exprimer mes sentiments les plus tendres à Valérie pour ses encouragements dans les derniers moments précédant la soutenance et pour tout ce qu'elle m'apporte.
Enfin, ma gratitude la plus chaleureuse est réservée à Guy, Carmen, Caroline, M. et Mme Cazade, ma famille de Saintes, d'Annecy et de l'Ain ainsi qu'à mes parents, à qui je dédie cette thèse, pour leurs précieux conseils et leur soutien sans limite.

Un grand merci ... à vous tous.

Table des matières

Table des figures

Introduction Générale

Le média audiovisuel tient une place de plus en plus importante dans la diffusion de l'information. En conséquence, il existe une masse sans cesse croissante d'informations audiovisuelles accessibles, se pose alors le problème de retrouver une information voulue dans cet ensemble. La numérisation et la création numérique de flux audiovisuels (AV) permettent leur exploitation dans des systèmes d'information audiovisuels. Cela nécessite une modélisation et une instrumentation des contenus des documents AV qui en autorisent l'accès direct et soient adaptées aux diverses utilisations possibles : recherche, indexation, navigation, etc...

Un document audiovisuel est représenté par la composition des flux audio et vidéo. Les méthodes de structuration de ces documents et d'indexation de leur contenu opèrent sur ces deux composantes. Ainsi, l'analyse et l'indexation de la vidéo peuvent s'effectuer à plusieurs niveaux de granularité : soit directement sur les pixels des images successives, soit dans l'espace des descripteurs préalablement extraits de la vidéo. Cette extraction requiert le développement d'outils d'analyse et de caractérisation du mouvement global (celui de la caméra), de segmentation des scènes animées en termes d'objets en mouvement, de détection des ruptures dans le flux temporel correspondant aux frontières des plans de montage, etc...

Les méthodes d'analyse s'appuient donc sur les outils méthodologiques classiques tels que la segmentation des images couleur, les méthodes d'estimation du mouvement de la caméra, mais aussi font intervenir les méthodes de décision statistique dans l'espace image ou l'espace des descripteurs.

La quantité croissante des contenus multimédia au format compressé pose des problèmes évidents quant aux choix de la représentation initiale : en bande de base (pixel), ou compressée (standards MPEG actuellement). Des méthodes sont proposées à ce jour pour ces deux domaines. Bien que dans plusieurs contextes, indexation et interprétation peuvent se limiter à un seul média (films muets, vidéo de télé-surveillance ou émissions de radio), dans la majorité des cas, il est judicieux de pouvoir analyser les deux composantes, vidéo et audio, ensemble. Les relations existantes entre ces deux flux peuvent donc être identifiées et donner lieu à une indexation plus sémantique. Étant entendu qu'il est possible de définir *l'indexation comme la recherche de repères temporels dans les flux audio et vidéo où certaines propriétés sont vérifiées*. Dans notre cas, la propriété que nous devons vérifier est la présence de frontières de scènes audiovisuelles dans les contenus multimédia télédiffusés.

De même, pour l'analyse et l'indexation audio, de nombreux travaux ont été entrepris dans ces domaines : analyse et synthèse de la parole, de la musique, etc... Ces méthodes et les résultats obtenus sont une source d'informations nécessaire pour une indexation de plus haut niveau sémantique dans le domaine du multimédia.

Les travaux entrepris ici sont focalisés sur l'indexation cross-média des contenus audiovisuels. L'analyse des contenus s'effectue dès lors non seulement sur l'étude du mouvement ou de mesure sur l'image vidéo mais aussi sur la combinaison de l'analyse des flux audio et vidéo. Ce n'est que lors de la dernière décennie que cette activité de recherche s'est considérablement développée. Dans cette branche de la recherche, les problèmes qui se posent concernent la fusion de l'information audio et vidéo. Cette dernière peut être opérée dans l'espace de décision où il s'agit de proposer les modèles de fusion des résultats des détecteurs multimédia. Par ailleurs, cette fusion peut être proposée dans l'espace commun des descripteurs.

Dans cette thèse, nous ne développons pas de méthodes pour toutes ces tâches. Il nous a semblé intéressant, dans un premier temps, de nous focaliser sur l'élaboration de modèles permettant de fusionner, dans l'espace de décision, les résultats d'indexation mono-média. Le tout en restant suffisamment générique et en s'appuyant sur les détecteurs bas-niveau tels que les changements de plan de montage (vidéo), silences (audio) ou changements du caractère de la bande audio. Étant entendu qu'une large gamme de méthodes ont déjà été proposées pour la détection des changements de plan, il nous semble plus opportun de se focaliser sur les problèmes de fusion même et d'analyse de la bande sonore.

Enfin, cette thèse s'est déroulée en partie en collaboration avec Philips Research Natlab (NL), dans le cadre du projet CASSANDRA[Cas]. Certaines méthodes d'analyse du média (détection de changement de plans et détecteur de silence) employées dans ce travail ont été développées par cette société et nous n'avons exploité que leurs résultats dans la démarche de fusion. Au contraire, les détecteurs de changements dans la bande sonore restent encore un vaste terrain de recherche, nous avons consacré une partie de cette thèse à l'étude des sons bruités et à leur segmentation. Le mémoire est organisé en quatre parties qui sont résumées par la suite :

Partie 1 : Introduction à l'analyse et à l'indexation audio et vidéo

Afin de replacer notre travail dans le contexte actuel de la recherche nous devons rappeler les travaux et méthodes développées dans ce domaine. En d'autres termes, nous considérons les méthodes d'analyse et d'indexation des flux audio et vidéo et nous rappelons en détail le mode opératoire ainsi que les algorithmes développés.

Dans un premier temps, nous présentons les méthodes d'analyse et d'indexation pour le flux vidéo.

Dans un second chapitre, nous mettons en avant les méthodes d'analyse et d'indexation des données sonores en introduisant les descripteurs et les méthodes d'analyse ad-hoc.

Enfin, nous terminons par un chapitre concernant les méthodes d'analyse cross-média.

Partie 2 : Fusion de l'information cross-média

La première partie des travaux de recherche menés au cours de ce travail de thèse ont été consacrés à la mise en œuvre d'une approche statistique de segmentation vidéo par l'analyse cross-média audio et vidéo du flux multimédia.

Dans ce chapitre, nous présentons un modèle de scène audiovisuelle.

Puis nous introduisons une statistique caractérisant les frontières de scènes et développons la règle de décision quant à la présence d'une frontière de scène par une approche bayésienne.

Partie 3 : Analyse et indexation des flux audio

Cette thèse a, entre autres, donné lieu à la mise en place d'une méthode d'analyse et d'indexation du flux audio.

Tout d'abord, nous présentons une méthode de détection des silences dans les flux audio compressés et non-compressés. Cette approche a été utilisée dans la méthode de caractérisation des frontières des scènes présentée dans la partie précédente.

Puis, nous nous intéressons à une méthode qui permet de détecter les transitions entre les bruits.

Enfin, nous introduisons une méthode de classification bayésienne des sons bruités pour sélectionner les descripteurs les plus adaptés au son bruité considéré.

Partie 4 : Expérimentations et résultats

L'ensemble des travaux présentés dans ce rapport de thèse ont amené à la réalisation de diverses expérimentations afin de valider les modèles présentés et de mettre en évidence les points forts et les points faibles des méthodes développées.

Dans un premier temps, nous présentons les résultats des expériences des méthodes de détection des silences dans le flux audio et de la détection des frontières des plans de montage dans le flux vidéo.

Puis, nous décrivons les expérimentations et présentons les résultats obtenus pour la méthode de détection de frontières des scènes audiovisuelles.

Ensuite, les résultats de la méthode de segmentation des bruits ainsi que la validation de l'ensemble des descripteurs proposés sont développés.

Enfin, nous terminons par la présentation des résultats de la classification des sons bruités naturels que nous avons obtenus. Nous présentons, par la même occasion, une méthode de caractérisation des transitions entre les bruits au sein d'une même classe. Nous avons qualifié cette approche de segmentation semi-aveugle car nous utilisons les résultats de la classification pour sélectionner un descripteur spécifique en fonction de la classe de bruit considérée.

Première partie

Introduction à la problématique et état de l'art dans les domaines de l'analyse et de l'indexation audio et vidéo

Chapitre 1

Problématique

Les problèmes liés à l'indexation et à l'archivage des documents vidéo évoluent avec la croissance des capacités de stockage et de traitement des informations numériques. Le travail d'indexation, même si l'intervention humaine restera certainement nécessaire pour extraire et organiser des informations d'un haut niveau sémantique, peut être notablement facilité si nous disposons d'outils bien adaptés.

Les documents montés représentent la majorité des documents audiovisuels. Cela signifie que les différentes prises de vues ont été assemblées lors de la phase de production ou post-production vidéo. La tâche d'indexation est à l'opposé de la production. Il s'agit d'identifier les frontières des segments créées lors du montage. Ce découpage a comme objectif de reconstruire le *story-board* du document. Deux grands niveaux peuvent être considérés. Le niveau de base qui représente le découpage en segments contigus, ce sont les plans de montage. Néanmoins, comme nous allons le voir, ce découpage génère un grand nombre de redondances. Le vrai défi consiste à proposer un découpage de plus haut niveau d'interprétation, un découpage en scènes audiovisuelles [NMBP+04]. Ces scènes peuvent regrouper plusieurs plans adjacents au cours du temps. Un exemple d'application d'un tel découpage est la recherche d'informations dans les archives audiovisuelles numérisées ou encore l'incorporation de ces techniques d'indexation dans les dispositifs numériques grand public de demain tels que les magnétoscopes, caméscopes, etc...

La segmentation en scènes audiovisuelles est une activité de recherche récente dans le domaine du multimédia. Elle suit l'évolution des moyens de télécommunication dont l'évolution implique la gestion de bases de données de plus en plus grandes. Ce nouveau challenge pose de nouveaux problèmes. En effet, tout d'abord la segmentation en scènes est directement dépendante de l'analyse des flux audio et vidéo puisqu'elle en utilise les résultats. De plus, il est aussi important de définir correctement un modèle de scène cohérent. La grande diversité des contenus audiovisuels actuels rend difficile la tâche précédente car il est a priori difficile de définir un modèle générique de segmentation en scènes. C'est la raison pour laquelle certains auteurs se sont focalisés sur un modèle de scène spécifique : dans les documents sportifs [KGG+03] ou bien les scènes de dialogue [AAW01].

Dans le premier chapitre, nous présentons les méthodes d'analyse et de segmentation dans le domaine de la vidéo. Dans un premier temps, les méthodes et algorithmes de segmentation

du flux vidéo en plans de montage sont présentés. Cette segmentation est de nos jours largement explorée dans la littérature, nous ne présentons que les principales techniques utilisées le plus couramment. Puis, dans la seconde partie de ce chapitre, nous nous intéressons aux modèles de scène proposés dans la littérature ainsi que les méthodes de segmentation liées à ces modèles.

Dans le second chapitre, nous traitons des méthodes d'analyse et d'indexation des flux audio. De la même manière que pour le chapitre dédié au domaine de la vidéo, nous avons découpé ce chapitre en deux grandes parties. D'une part, nous présentons les propriétés des sons et la manière dont le système auditif humain perçoit les signaux audio. Puis, nous présentons les principaux descripteurs audio ainsi que les méthodes d'analyse proposées dans la littérature. Celles-ci permettent, respectivement, de modéliser les propriétés perceptives des signaux sonores et de préparer le terrain pour les méthodes d'indexation audio : classification audio [TC02], segmentation en locuteurs [RR95], etc... Dans la deuxième partie de ce chapitre, nous nous intéressons aux diverses méthodes d'indexation des flux audio. Enfin, une étude des travaux d'indexation des contenus audiovisuels est présentée dans le troisième chapitre de cette partie. Ces méthodes consistent à fusionner les informations issues de l'analyse conjointe des flux audio et vidéo.

Chapitre 2

État de l'art dans le domaine de l'indexation vidéo

Dans ce chapitre, nous présentons les travaux qui ont été menés dans le domaine de l'indexation vidéo par le contenu. Cette dernière est devenue essentielle compte tenu de l'évolution des techniques de compression audio et vidéo ainsi que celle des supports de stockage. Ainsi, de nos jours, l'utilisateur doit pouvoir gérer facilement une quantité de données vidéo toujours plus importante proportionnellement à l'évolution des capacités des supports de stockage.

Nous précisons que dans cette analyse nous nous intéressons aux méthodes utilisant des indices globaux pour une indexation temporelle des documents vidéo. Une vaste fouille des méthodes destinées à une indexation plus sémantique (intérieur/extérieur, objets) reste en dehors de cette étude.

Les contenus vidéo de type artistique, par opposition aux contenus vidéo acquis en direct, ont subi une phase de post-traitement : le montage vidéo. Le but principal de l'indexation vidéo consiste à retrouver, par des méthodes automatiques, la structure temporelle des vidéos conçue au cours de la phase de montage. De plus, cette structuration est hiérarchique [AJ94] (cf Fig. 1).

Enfin, il est important de préciser que les applications liées à l'indexation vidéo, comme la navigation rapide dans les bases de données vidéo ou bien encore la génération de résumés vidéo, ne peuvent pas se faire sans avoir au préalable retrouver la structuration temporelle hiérarchique d'un document vidéo. Ce qui signifie que l'indexation vidéo est une chaîne d'opérations dont les premiers maillons sont constitués par la détection des plans de montage puis des scènes.

2.1 Partitionnement en plans de montage

Cette phase de partitionnement temporel des documents vidéo en plans de montage constitue l'étape initiale des traitements liés à l'analyse des contenus vidéo. Dès lors, le plan est l'entité élémentaire de la représentation d'un document multimédia après les images elles-mêmes.

La détection des frontières des plans de montage consiste à retrouver les frontières de collage qui ont été faites lors de la phase de montage. Il existe différents types de transitions

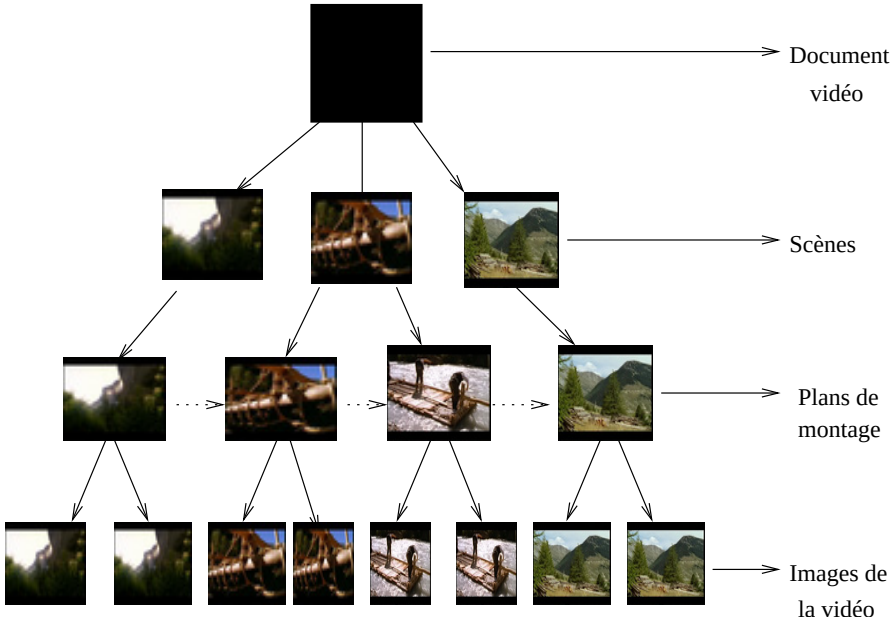


FIG. 1 – Structuration hiérarchique d'un document vidéo artistique

vidéo entre deux plans de montages mais deux grandes familles se distinguent :

- les transitions simples, ou abruptes, caractérisant un changement radical du contenu : les *cuts*.
- les transitions progressives, ou graduelles, caractérisant un changement progressif du contenu : les *fondus*, *volets* et autres.

Depuis de nombreuses années, diverses solutions ont été proposées au problème de l'identification des frontières des plans de montage [EB04, LSL97, MF03, CBPLB99, PR01, She97]. Certaines techniques mises en oeuvre pour la détection des cuts sont basées sur un calcul de distance inter-histogrammes de couleur sur deux images successives [SHM+03, She97], ou sur des sous-parties des images successives [PR01, MF03]. D'autres méthodes, plus complexes, se basent sur l'estimation du mouvement de la caméra via l'estimation du flot optique [BP02, BGG99, CBPBM99, CBPLB99, NPZ02, KD97, TSKR00, DBP01] ou bien sur le suivi des objets en mouvement [CSZK01]. En ce qui concerne la détection des transitions vidéo progressives telles que les volets, les fondus enchaînés ou bien encore les balayages, des études spécifiques ont été menées et reposent sur une modélisation ad-hoc de chacune de ces transitions [JYL00]. Dans [BGR+99], l'auteur propose une approche basée sur l'estimation du mouvement de la caméra représentée par un modèle affine 2D pour détecter à la fois les cuts et les transitions progressives.

Finalement, nous pouvons en déduire que deux grandes approches principales émergent pour résoudre le problème de la détection des transitions vidéo :

- les approches basées sur l'étude d'histogrammes et
- les méthodes utilisant l'estimation du mouvement de la caméra par le flot optique, utilisées dans le cas de flux vidéo compressés ou en bande de base (non-compressé).

À cela viennent s'ajouter des méthodes basées sur des réseaux de neurones ou bien des méthodes statistiques. Toutefois, nous ne traiterons pas de ces méthodes car ces dernières sont

utilisées, le plus souvent, pour la détection des transitions graduelles plus difficilement détectables et qui font appel à des approches plus complexes.

Enfin, nous remarquons que dans la campagne internationale d'évaluation des méthodes d'indexation TRECVID[1] [WIB+03] la tâche de découpage en plans de montage requiert l'identification en deux classes d'effets : cuts et progressifs.

2.1.1 Détection des frontières de plans de montage par études d'histogrammes

Ces méthodes basées sur des calculs à partir des histogrammes de couleurs ont été, avec les méthodes basiques de différence entre pixels[ZK93], les premières approches mises au point pour la détection des transitions vidéo. Toutes ces méthodes ont la même base, à savoir l'extraction et le calcul des histogrammes de couleurs à partir des images extraites de la vidéo non compressée. En revanche, les méthodes mises en œuvre pour l'analyse de cette extraction varient suivant les auteurs. Dans les méthodes les plus simples, la décision concernant la détection de la limite d'un plan est prise suivant un seuillage sur la différence des histogrammes entre deux images consécutives. Cette différence est calculée de la manière suivante :

$$\Delta(t) = \sum_{j=0}^{a} |h_t(j) - h_{t-1}(j)| \tag{1}$$

avec h_t l'histogramme de la t-ième image et a le nombre de casiers de l'histogramme.

Pickering et Rüger [PR01]

Dans [PR01] les auteurs proposent une méthode de détection des frontières des plans de montage appliquée aux corpus vidéo fournis lors de la campagne d'évaluation TREC-10. Cette approche est basée sur l'extraction des histogrammes dans l'espace de couleur RGB. Chaque image de la vidéo est divisée en neuf blocs de taille identique, et pour chaque bloc les auteurs déterminent un histogramme pour chacune des composantes de l'espace de couleur RGB. Après quoi, la distance *Manhattan* est calculée pour chaque composante des histogrammes relativement à deux blocs correspondants dans deux images distinctes dans la vidéo. La plus grande distance parmi les trois composantes RGB d'un bloc est retenue comme distance associée à ce bloc. Enfin, après avoir calculé les distances pour chacun des blocs, la distance entre deux images est caractérisée par la distance médiane parmi les neuf distances de blocs. Cette mesure peut être écrite de la manière suivante :

$$d_n(t) = \frac{1}{n} \sum_{i=0}^{n-1} D(t+i, t-n+i) \tag{2}$$

où $D(i,j)$ représente la distance médiane des blocs entre les images i et j.

Les auteurs admettent comme pic pour d_n toute valeur à la fois supérieure à un seuil pré-défini et supérieure tant aux 16 valeurs de d_n qui la précèdent qu'aux 16 qui la succèdent. Ainsi, une transition est détectée s'il y a un pic qui correspond à d_{16} ou bien à d_8. S'il y a la

[1] http ://www-nlpir.nist.gov/projects/t01v/

présence d'un pic supplémentaire correspondant à d_2 alors la transition est étiquetée en tant que *cut*, sinon la transition est considérée comme *graduelle*. Bien que l'algorithme présenté ici soit relativement fiable quant à la détection des transitions en tant que telles, il ne permet pas de déterminer les numéros d'image de début et de fin des transitions graduelles. Pour tenter d'y remédier les auteurs proposent un algorithme pour donner une estimation des bornes de l'intervalle pour les transitions graduelles : à chaque image, la différence d_4 est comparée au seuil pré-défini. Si cette valeur est plus grande que le seuil, alors l'image correspondante est considérée comme potentielle image de début de la transition. Toutefois, cette décision reste provisoire car si, après examen des images suivantes, la valeur d_4 devient inférieure au seuil avant que la transition ne soit détectée alors le potentiel début de transition précédemment estimé est éliminé et la recherche continue. L'image de fin de transition correspond au moment où d_4 retombe pour la première fois en-dessous du seuil juste après la détection de la transition.

La méthode proposée dans cet article affiche des résultats très corrects pour la détection des cuts : Rappel = 93,20% et Précision = 91.60%. En revanche, ces derniers sont plutôt moyens dans le cas des transitions graduelles : Rappel = 64,10% et Précision = 67.40%.

Shen [She97]

Dans [She97], l'auteur présente un algorithme de détection des cuts dans les vidéos MPEG compressées faisant appel aux histogrammes et à la distance de Hausdorff à des échelles différentes au niveau des images de contour. La distance de Hausdorff est une mesure de distance entre deux ensembles de points d'une image. Pour chaque point des deux ensembles, en premier lieu il faut déterminer la plus petite distance à tous les autres points de l'autre ensemble. Enfin, la distance Hausdorff est la plus grande valeur parmi les plus petites distances précédemment calculées. Dans l'article, les auteurs extraient des points de contour des images de type I dans les vidéos de type MPEG puis proposent un algorithme de détection des cuts. L'approche proposée admet comme hypothèse que les contours présents dans une petite région de l'image ne doivent pas changer ou doivent présenter un mouvement similaire si l'on compare deux images successives d'un même plan. De plus, le mouvement des points de contour de régions adjacentes doit être relativement similaire. En revanche, la distribution des contours doit changer de façon significative en présence d'une coupure.

Dans cet article, Shen utilise une généralisation de la distance de Hausdorff pour comparer la distribution des pixels de contour entre les blocs de deux images successives. Cette généralisation s'appelle la meilleure distance partielle et est définie comme suit :

$$h_K(A,B) = K^{th}_{a\in A} min_{b\in B}||a-b|| \tag{3}$$

où A, B sont deux ensembles de points et K est la $k-$ième meilleure distance.

Dans le cas particulier de la distance de Hausdorff ($K = 1$) l'équation 3 devient :

$$h(A,B) = max_{a\in A} min_{b\in B}||a-b|| \tag{4}$$

Ainsi, l'auteur fixe la valeur maximale de h_k à h et calcule le nombre K de pixels de A dont la distance à B est inférieure à h. Pour chaque translation des pixels de contour du bloc, nous obtenons une valeur de K et l'histogramme $K_{i,j}$ de déplacement i, j est calculé par :

$$k_{i,j} = \sum_{a_{x,y}\in A} f_{i,j}(a_{x,y}), |i| \leq D_x \text{ et } |j| \leq D_y \tag{5}$$

où $f_{i,j}(a_{x,y}) = 1$ si $a_{x,y} \in A$ et $b_{x+i,y+j} \in B^+$ avec B^+ l'image dilatée d'un disque de largeur h. D_x et D_y sont les distances maximales permises définissant ainsi une fenêtre dans laquelle on calcule l'appariement.
La valeur de h dépend de la tolérance admise par rapport à l'erreur de position des pixels de contour calculés et au mouvement relatif admis. Ces histogrammes sont calculés sur des blocs 8×8 de l'image. Afin que la détection soit plus robuste aux mouvements et au bruit, la méthode regroupe les blocs 4×4 faisant la somme des histogrammes échantillonnés de chaque bloc. L'échantillonnage consiste à ne garder que les valeurs importantes des histogrammes calculés sur des fenêtre superposées à l'intérieur de chaque bloc. On suppose que les valeurs les plus faibles correspondent au bruit ou à des mouvements incohérents. Après 4 phases successives de regroupement des blocs 4×4 un seul histogramme est obtenu. Finalement, la décision est prise par rapport à ce dernier histogramme après normalisation par rapport au nombre de contours dans l'image. L'algorithme de détection recherche d'éventuels pics dans l'histogramme, en effet un mouvement de la caméra est caractérisé par un fort pic au niveau de l'histogramme. En revanche, une transition dans la vidéo est matérialisée par un pic plus faible au niveau de l'histogramme.

Pour conclure, les performances affichées par cette approche sont très élevées comprises entre 92 et 100 % de rappel et entre 91 et 100 % pour la précision mais les tests ont été effectués sur un corpus vidéo composé de 28775 images, donc relativement restreint.

Mas et Fernandez [MF03]

Les auteurs, dans [MF03], proposent un algorithme de détection des frontières des plans de montage dans le cadre de la campagne d'évaluation TRECVID 2003. Cette méthode est basée sur un algorithme de détection multi-échelle dans lequel les différences des histogrammes de couleurs sont considérées et se rapprochent de la méthode décrite dans [PR01]. Les auteurs montrent que la différence entre les histogrammes est une métrique plus appropriée que la différence cumulée inter-image (voir équation (6)). Cette dernière est très sensible au mouvement de la caméra. Cette mesure est définie par :

$$g(n, n+k) = \sum_{x,y} |I_n(x,y) - I_{n+k}(x,y)| \tag{6}$$

où $I_n(x,y)$ correspond au niveau d'intensité de l'image n au point de coordonnées x,y et $g(n, n+k)$ est la valeur de similarité entre les images n et $n+k$ avec $k \geq 1$.
C'est la raison pour laquelle les auteurs plaident en faveur des histogrammes couleur dans l'espace couleur RGB pour obtenir de meilleures performances de détection. Dans cet article, les auteurs proposent une mesure de différence des histogrammes de couleur pour la détection des transitions vidéo. Cette mesure est différente de celle présentée dans [PR01] et s'écrit sous la forme suivante :

$$d_{RGB}(X,Y) = \sum_{i=1}^{M} |h_x(i) - h_y(i)| \tag{7}$$

avec h_x correspondant à l'histogramme couleur de l'image X contenant M casiers.

La prise de décision pour la détection des frontières des plans de montage est basée sur un

seuillage des valeurs de la fonction suivante :

$$HistDif[i] = d_{RGB}(j, j-1) \sum_{i=1}^{M} |h_j(i) - h_{j-1}(i)| \tag{8}$$

où h_j représente l'histogramme couleur sur M casiers de l'image numéro j dans la vidéo. Le graphe de la fonction $HistDif$ au cours du temps produit un pic lors de la présence d'un cut. Alors qu'une transition graduelle est matérialisée par une variation plus douce avec une certaine durée. Cependant, pour améliorer cette détection, les auteurs ont décidé de convoluer le signal produit par la fonction $HistDif$ avec une fenêtre de taille W :

$$HistDif_{conf}[j] = HistDif[j] * \frac{1}{W} \times rect\left(\frac{j}{W}\right) \tag{9}$$

avec $rect(x) = \begin{cases} 1, |x| \geq 1/2 \\ 0, \text{sinon} \end{cases}$

L'intérêt d'une telle convolution est de lisser le signal afin d'éliminer les hautes fréquences sources de fausses alarmes lors de la phase d'analyse du signal en vue de la détection des transitions. Enfin, le dernier problème est de pouvoir discerner entre un pic abrupt, synonyme de la présence d'un cut, et une variation plus douce, synonyme d'une transition graduelle. Dans le cas des cuts, les auteurs proposent une méthode qui consiste à prendre la première dérivée du signal convolué (voir formule 9) :

$$HistDif^{deriv}_{conv}[j] = [1, -1] \times HistDif_{conv}[j] \tag{10}$$

Ainsi les pics présents dans le signal convolué seront représentés comme une paire de pics l'un positif et l'autre négatif dans un intervalle de temps très court.

En ce qui concerne les transitions graduelles, les auteurs proposent une méthode plus complexe consistant à appliquer au signal précédemment convolué (voir formule 9) des opérateurs morphologiques. Ces opérateurs sont liés à des notions telles que l'*érosion* et la *dilatation* définies respectivement par les formules suivantes :

$$\epsilon_B(f(x)) = inf_{y \in B}[f(x-y)] \tag{11}$$

$$\delta_B(f(x)) = sup_{y \in B}[f(x-y)] \tag{12}$$

où B est un ensemble de points.

Malgré les algorithmes mis en œuvre, les résultats restent moyens quant à la détection des transitions graduelles avec un rappel aux alentours de 50% et une précision de 80%. En ce qui concerne la détection des cuts, le rappel est de l'ordre de 95% et la précision de 75% sur le corpus vidéo de TRECVID 2003.

2.1.2 Caractérisation des ruptures de plans de montage par l'estimation du mouvement de la caméra

Coudert, Benois-Pineau, Le Lann et D. Barba [CBPLB99]

Dans [CBPLB99], les auteurs proposent un enchaînement de méthodes permettant dans un premier temps d'estimer le mouvement global de la caméra modélisé par un modèle affine

1D dans le domaine de la transformée "Mojette", transformée discrète de Radon. Puis, en utilisant les résultats de l'analyse précédente, le système offre la possibilité de la détection des frontières des plans de montage dans les contenus compressés. Enfin, l'approche présentée permet d'extraire et de suivre les zones d'intérêt, à savoir les objets en mouvement, à l'intérieur des plans de montage précédemment détectés.

Le système proposé calcule, pour tous les couples d'images successives, la transformée de la Mojette définie comme la transformée discrète de Radon. La transformée de Radon est définie par :

$$R_{\phi}[I](u) = \int\int_{D} I(x,y)\delta(u - x.sin(\phi) - y.cos(\phi))dxdy \tag{13}$$

avec $I(x,y)$ l'intensité lumineuse du point de coordonnées (x,y) dans l'image I, δ la fonction de Dirac et ϕ l'angle de projection.
Dans la cas discret, les pixels de l'image sont modélisés par un modèle de Dirac et une direction de projection $tan(\phi) = -\frac{p}{q}$ avec p et q premiers. D'où la formulation suivante :

$$M_{p,q}[I](m) = \sum_{k,l} I(k,l)\delta(m + q \times k - p \times l) \tag{14}$$

avec $M(m)$ le $m-$ième élément de la transformée de Mojette, $I(k,l)$ l'intensité lumineuse du point de coordonnées (k,l) dans l'image I et $m + q \times k - p \times l$ représente la droite définie par la direction de projection et la valeur de m.

Dans le cadre des travaux, les auteurs ont opté pour le modèle de Haar comme modèle de pixel d'où l'équation suivante :

$$Mh_{p,q}(m) = \sum_{t} M(m).C_{p,q}(t - m) \tag{15}$$

avec C une fonction linéaire définie par morceaux.

L'estimateur de mouvement récupère les données issues de la transformée de Mojette. Ici, il s'agit d'estimer les paramètres du modèle 1D du mouvement. Pour cela, les auteurs proposent un estimateur robuste basé sur le calcul d'un ensemble de mesures de similarités locales. Une mesure locale de similarité correspondant au coefficient de corrélation calculé pour chaque angle ϕ de projection de deux images successives est définie par :

$$\rho_j(m) = \frac{cov_k[M_j^t(m), M_j^{t+1}(m + dm_j)]}{sqrt(var_k[M_j^t(m)].M_j^{t+1}(m + dm_j))} \tag{16}$$

avec K la taille de la fenêtre d'analyse, $M(m + dm_j)$ l'interpolation linéaire de la projection selon le vecteur de déplacement dm_j et cov et var respectivement les fonctions de covariance et de variance classiques.
À partir de l'étude des valeurs prises par ces coefficients de corrélation ρ_j, les auteurs ont remarqué que dans le cas d'un mouvement continu et de non changement de plan de montage les valeurs de ρ_j étaient proches de 1. En revanche, dans le cas de changement de plan de montage ou de la présence d'un mouvement plus complexe les valeurs des coefficients de corrélation chutent. C'est la raison pour laquelle les auteurs proposent un algorithme robuste de détection des frontières des plans de montage basé sur les informations fournies par l'estimateur de mouvement décrit précédemment. Néanmoins, les valeurs ρ sont trop bruitées et

donc peu exploitables c'est pourquoi il est proposé de considérer les valeurs moyennes, $\overline{\rho}$. Ces valeurs permettent de lisser la courbe et donc d'éliminer les hautes fréquences synonymes de fausses alarmes. Enfin, le test de Hinkley [Bas88] est utilisé pour étudier les variations de la courbe des valeurs $\overline{\rho}$ afin de déterminer la présence ou non d'une transition vidéo quelle soit de type cut ou bien de type graduelle.

Les transitions de type cut sont caractérisées par un pic soudain dans un intervalle de temps très court alors que les transitions progressives sont matérialisées par une évolution plus douce avec une certaine durée. Enfin, le dernier maillon de la chaîne d'analyse développée dans ce papier est l'extraction et le suivi des zones d'intérêt à l'intérieur même de chaque plan de montage préalablement détecté. Les zones d'intérêt peuvent être assimilées à des sous-parties d'une image contenant un objet en mouvement. Pour cela, les auteurs se basent sur l'analyse des valeurs de ρ. Les valeurs de ρ dites *rejetées*, liées aux vecteurs *outliers* obtenus lors du processus d'estimation du mouvement global de la caméra, matérialisent les zones d'intérêt.

Les résultats ont été menés sur un corpus vidéo fourni par l'INA. Les résultats obtenus sont très prometteurs du fait que toutes les transitions présentes dans la vidéo ont été détectées, seules les marques de début et de fin de transitions graduelles sont plus approximatives. Toutefois, les auteurs n'ont pas évalué leur algorithme de détection sur un panel plus vaste de vidéos contenant différents genres comme les journaux télévisés, ou bien les clips vidéos par exemple. Le seul type de vidéo traité est le documentaire.

Bruno et Pellerin [BP02]

E. Bruno et D. Pellerin présentent dans [BP02] une méthode de détection des transitions vidéo basée sur la prédiction linéaire du mouvement. La méthode proposée consiste à estimer les coefficients de la transformée en ondelettes du flot optique. Pour commencer, les auteurs rappellent l'hypothèse fondamentale de la conservation de l'intensité lumineuse selon laquelle l'intensité lumineuse d'une image $I(p_i, t+1)$ reste inchangée le long du déplacement entre t et $t+1$:

$$I(p_i, t) = I(p_i + v(p_i), t+1) \tag{17}$$

avec $v(p_i, t) = (u, v)$ le vecteur vitesse au pixel p_i entre les images $I(p_i, t)$ et $I(p_i, t+1)$.

À partir de cela, les auteurs proposent de modéliser le champ de mouvement par un développement hiérarchique en séries d'ondelettes à 2 dimensions de l'échelle L à l :

$$\begin{aligned} v_\Theta(p_i) = & \sum_{k_1,k_2=0}^{2^L-1} c_{L,k_1,k_2} \Phi_{L,k_1,k_2}(P_i) \\ & + \sum_{J \geq L}^{l} \sum_{k_1,k_2=0}^{2^j-1} [d^H_{j,k_1,k_2} \Psi^H_{j,k_1,k_2}(p_i) \\ & + d^D_{j,k_1,k_2} \Psi^D_{j,k_1,k_2}(p_i) + d^V_{j,k_1,k_2} \Psi^V_{j,k_1,k_2}(p_i)] \end{aligned} \tag{18}$$

avec $\Phi_{L,k_1,k_2}(P_i)$ la *fonction d'échelle* et $\Psi^{H,D,V}_{j,k_1,k_2}(p_i)$ les *ondelettes* respectivement horizontales, diagonales et verticales du flot optique. Le modèle est multi-échelle, et s'applique du niveau d'échelle le plus large L au niveau d'échelle le plus fin l.
Le vecteur des paramètres de mouvement, Θ, contenant les coefficients des ondelettes c_{L,k_1,k_2}

et $d^{H,D,V}_{j,k_1,k_2}$ pour tous les j, k_1, k_2 est estimé par la minimisation d'une fonction objective :

$$\Theta = argmin_{\Theta} \sum_{p_i \in \Omega} \rho(I(P_i + v_{\Theta}(P_i), t+1) - I(p_i, t)) \tag{19}$$

avec $\rho(.)$ un M-estimateur robuste.

Le processus de minimisation utilisé dans cet article est celui mis au point dans [OB95]. Après avoir défini un modèle d'estimation du flot d'optique basé sur la décomposition en séries d'ondelettes, les auteurs proposent dès lors d'appliquer cette estimation à la détection des transitions vidéo. Pour cela, une estimation grossière est suffisante pour pouvoir distinguer si une variation temporelle de l'image est due au mouvement ou bien à la présence d'une transition. C'est pourquoi, les auteurs fixent l à 3 dans l'équation 18 et par conséquent la dimension du vecteur des paramètres de mouvement Θ est égale à 128. Plutôt que d'étudier les variations des normes des vecteurs des coefficients d'ondelettes Θ, pas suffisamment discriminants car $||\Theta||$ peut prendre une amplitude similaire qu'il s'agisse d'un fort mouvement ou bien de la présence d'une transition, les auteurs font appel à la méthode de prédiction linéaire d'erreur de Θ. Le principe consiste à prédire une valeur x_n, d'un signal stationnaire $\{x_1, x_2, \ldots, x_N\}$, en se basant sur les valeurs, x_{n-i}, précédentes :

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + e_n \tag{20}$$

avec p l'ordre de la prédiction, a_i les coefficients de prédiction et e_n l'erreur de prédiction. Les coefficients a_i sont estimés par la minimisation de l'erreur de prédiction quadratique e_n^2. Ici, les auteurs proposent d'appliquer le principe de la prédiction linéaire aux coefficients du vecteur Θ. D'où $e_n = (e_{1n}, \ldots, e_{ln})$ est le vecteur de prédiction d'erreur contenant les coefficients de prédiction linéaire estimés à partir des coefficients d'ondelettes :

$$e_n = \Theta_n - \sum_{i=1} p a_i \Theta_{n-i} \tag{21}$$

La norme de e_n est grande dans le cas d'une discontinuité dans les coefficients d'ondelettes et faible sinon. Ainsi, il est maintenant plus facile de détecter les transitions vidéo puisqu'il ne reste plus qu'à détecter les pics de la courbe des valeurs de e_n.

Les résultats sont présentés en fonction du genre vidéo considéré. Les auteurs ont utilisé deux genres vidéo : une vidéo d'un journal télévisé d'une durée d'environ 4 minutes et une vidéo d'une série d'une durée de 3 minutes. Pour le journal télévisé le rappel est de 90% et la précision de 100%. Quant à la série le rappel est de 100% et la précision aux alentours de 99%. Cependant, ces valeurs restent à confirmer compte tenu de la faible quantité de données présentes dans le corpus vidéo utilisé. Néanmoins, la détection des ruptures basée sur l'estimation du mouvement seule n'est pas suffisante. C'est pourquoi, les indices spectraux doivent être pris en compte.

Bouthemy et Ganansia [BG96]

Dans [BG96] les auteurs proposent une approche pour la détection des transitions vidéo et la caractérisation du mouvement de la caméra basée sur l'estimation du flot optique. Cette

même approche est complétée dans [BGG99]. Les auteurs précisent que l'estimation du mouvement de la caméra par un modèle affine 2D est suffisant pour la détection des transitions vidéo mais pas assez pour pouvoir caractériser les mouvements de la caméra : zoom, fondu, etc... Pour cela, il faut utiliser un modèle 3D du mouvement de la caméra. La méthode d'estimation du mouvement de la caméra s'appelle RMR (Robuste et Multi-Résolution)[OB95]. Cette méthode est basée sur la minimisation d'un critère, le M-estimateur, pour assurer une estimation robuste en cas de mouvements secondaires ou bien dans des zones de l'image où l'équation classique du mouvement n'est pas valide. Les auteurs utilisent un modèle affine du mouvement à 6 paramètres, w_Θ. Ce modèle est défini, pour un point $p(x, y)$ par rapport au point de référence (x_g, y_g), par l'équation suivante :

$$w_\Theta(p) = \begin{pmatrix} a_1 + a_2(x - x_g) + a_3(y - y_g) \\ a_4 + a_5(x - x_g) + a_6(y - y_g) \end{pmatrix} \tag{22}$$

avec $\Theta = (a_1, a_2, a_3, a_4, a_5, a_6)^T$ le vecteur des paramètres de mouvement à estimer.
L'estimation des éléments de Θ s'effectue entre deux images successives $I(t)$ et $I(t+1)$ comme suit :

$$\widehat{\Theta} = argmin_\Theta \sum_{p_i \in I} \rho(DFD_\Theta(p_i)) \tag{23}$$

avec $DFD_\Theta(p_i) = I(p_i + \vec{w}_\Theta(p_i), t + 1) - I(p_i, t)$, $I(p, t)$ l'intensité du pixel au point p sur l'image I à l'instant t et $\rho(x)$ le M-estimateur robuste.

Dans cet article, les auteurs utilisent la fonction de Tukey comme estimateur. Cette fonction $\rho(x)$ et sa dérivée $\Psi(x)$ sont définies respectivement par :

$$\rho(x, C) = \begin{cases} \frac{x^6}{6} - \frac{C^2 x^4}{2} + \frac{C^4 x^2}{2} & \text{si } |x| < C \\ \frac{C^6}{6} & \text{sinon.} \end{cases} \tag{24}$$

$$\Psi(x, C) = \begin{cases} x(x^2 - C^2)^2 & \text{si } |x| < C \\ 0 & \text{sinon.} \end{cases} \tag{25}$$

avec C un paramètre.
La fonction de Tukey a été choisie au détriment de l'estimateur des moindres carrés pour sa meilleure robustesse. L'estimation est obtenue par des étapes successives où les paramètres du mouvement sont estimés à différentes résolutions de l'image. Ces différentes résolutions sont obtenues par filtrage gaussien. En effet, à chaque instant la méthode construit la pyramide gaussienne de l'image. Au niveau de la plus basse résolution de l'image, nous avons à estimer le vecteur des paramètres de mouvement suivant :

$$\widehat{\Theta^0} = argmin_{\Theta^0} \sum_{p_i \in I} \rho(r_i) r_i = I(p_i, t + 1) - I(p_i, t) + \nabla I(p_i, t).\vec{w}_{Theta^0}(p_i) \tag{26}$$

avec r_i la différentielle de premier ordre de la fonction $DFD_\Theta(p_i)$ et ∇I le gradient spatial de l'intensité lumineuse pour l'image I au niveau de la résolution la plus basse.
Les résolutions successives sont obtenues à partir de l'estimation dans la résolution antérieure :

$$\widehat{\Theta}^{k+1} = \widehat{\Theta}^k + \nabla \widehat{\Theta}^k \tag{27}$$

La méthode itère le calcul de l'estimation de Θ (voir 27) jusqu'à tendre vers un critère de convergence ou bien encore la méthode s'arrête lorsqu'un nombre d'itérations spécifié a été atteint.

Les auteurs proposent d'appliquer les résultats précédents au problème de détection des transitions vidéo. Pour cela, les auteurs rappellent qu'à chaque étape de l'estimation, donc pour chaque résolution, le problème de la minimisation (voir équation 26) est équivalent à :

$$\widehat{\Delta\Theta} = argmin_{\Delta\Theta} \sum_{p_i} \frac{1}{2} w_t r_i^2 \tag{28}$$

avec $w_i = \frac{\Psi(r_i)}{r_i}$ indiquent si un point appartient à la région de l'image qui suit le mouvement global, Ψ la fonction dérivée de la fonction ρ et r_i défini par l'équation 26.
Cette méthode équivalente s'appelle IRLS(Iteratively Re-weighted Least Squares).

De plus, les auteurs définissent un support, S_d, du mouvement global dominant comme un ensemble de points p_i satisfaisant $w_i > \nu$, où ν est un seuil fixé (ici 0.2). Le nombre de pixels faisant partie du support, n_d, reste constant au sein d'un même plan mais chute brusquement au voisinage de 0 dans le cas contraire. Dans le cas d'une transition graduelle, n_d décroît moins fortement que dans le cas d'une transition abrupte ce qui permet de faire la distinction entre les deux types de transitions. Concernant la détection des transitions, les auteurs partent du principe qu'au sein d'un même plan les supports qui correspondent au mouvement global compris entre $(t-1,t), S_0$ et $(t,t+1), S_d$ doivent être proches. C'est pourquoi la décision concernant la détection d'une frontière de plan est prise en fonction de la valeur prise par la mesure $\phi_t = \frac{n_d}{n_0}$. Pour détecter les variations de la courbe des valeurs de ϕ_t, les auteurs font appel au test de Hinkley défini comme :

$$\begin{cases} S_k = \sum_{t=0}^{k} (\zeta_t - m_0 + \frac{\delta_{min}}{2}) \quad (k \geq 0) \\ M_k = max_{0 \leq i \leq k} S_i \text{ détection si } M_k - S_k > \alpha \end{cases} \tag{29}$$

$$\begin{cases} T_k = \sum_{t=0}^{k} (\zeta_t - m_0 - \frac{\delta_{min}}{2}) \quad (k \geq 0) \\ N_k = min_{0 \leq i \leq k} T_i \text{ détection si } T_k - N_k > \alpha \end{cases} \tag{30}$$

avec m_0 la moyenne des ϕ_t, δ_{min} un saut négatif ou positif que l'on cherche à détecter et α un seuil fixé.
Cette méthode permet de détecter les transitions de type cut, fondus et volets.

Au niveau des résultats, les tests ont été menés sur un grand nombre d'extraits vidéo au format MPEG2. Les auteurs ne donnent pas leur résultats sous forme de rappel et précision mais donnent une estimation du temps de calcul de la méthode qu'ils proposent dans cet article. Pour un couple d'image de taille 256×256 pixels, il faut compter $1.5s$ sur une station de travail *Sun-UltraSparc* et seulement $0.9s$ si l'on ne calcule pas l'estimation de mouvement et que l'on utilise les vecteurs de mouvement déjà présents dans le flux MPEG. Pour l'évaluation de la méthode de détection des transitions vidéo, les auteurs ont fixé $\alpha = 0.1$ et $\delta_{min} = 0.2$ dans les équations 29 et 30. Une partie des résultats est présentée sous forme de courbes mais les auteurs ne nous renseignent pas sur la durée exacte de leur corpus de test ainsi que sur les résultats globaux de la méthode, avec et sans le calcul de l'estimation de mouvement, sur l'ensemble de la base de test vidéo.

Durik et Benois-Pineau [DBP01]

Concernant les méthodes d'estimation robustes du mouvement de la caméra appliquées à la détection des transitions vidéo dans les flux vidéo compressés MPEG2, nous allons terminer

cet état de l'art par l'approche présentée dans [DBP01].
Dans cet article, les auteurs proposent une méthode d'estimation du mouvement basée sur l'estimation robuste des 6 paramètres de mouvement du modèle affine 3D. Plus précisément les auteurs proposent de comparer le modèle affine à 3 paramètres, plus couramment utilisé, avec le modèle affine à 6 paramètres plus complexe mais plus robuste aux mouvements complexes de la caméra. Le modèle affine à 6 paramètres est défini comme :

$$\begin{cases} dx_i = a_1 + a_2 x_i + a_3 y_i \\ dy_i = a_4 + a_5 x_i + a_6 y_i \end{cases} \tag{31}$$

avec (x_i, y_i) la position du centre du $i^{ième}$ macro-bloc dans l'image courante, (dx_i, dy_i) le vecteur de mouvement MPEG2 pointant de la position courante vers la position de ce macro-bloc dans l'image précédente, uniquement dans le cas des images P et $\Theta = (a_1, a_2, a_3, a_4, a_5, a_6)^T$ le vecteur des paramètres de mouvements à estimer.

Selon les valeurs estimées pour ces paramètres, il est alors possible d'en déduire le type de mouvement de la caméra selon les correspondances suivantes :

$$\begin{array}{cc} pan = a_1 & tilt = a_4 \\ div = (a_2 + a_6)/2 & rot = (a_5 - a_3)/2 \\ hyp_1 = (a_2 - a_6)/2 & hyp_2 = (a_3 + a_5)/2 \end{array}$$

avec *pan* pour panoramique, *tilt* pour pente ou inclinaison, *rot* pour rotation et $hyp_{1,2}$ pour hyperbolique.
Dans le cas d'un mouvement de la caméra purement panoramique (respectivement incliné), seul le paramètre a_1 (respectivement a_4) correspondant au panoramique (respectivement à une inclinaison) n'est pas nul, tous les autres étant égaux à zéro. Dans les cas d'une caméra fixe, tous les paramètres sont nuls.

Le second modèle de mouvement présenté dans cet article est le modèle à 3 paramètres proposé dans [TSKR00]. Ce modèle est défini par :

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} p_1 & 0 \\ 0 & p_1 \end{pmatrix} \begin{pmatrix} x_i^{'} \\ y_i^{'} \end{pmatrix} + \begin{pmatrix} p_3 \\ p_4 \end{pmatrix} \tag{32}$$

avec (x, y) les coordonnées du point dans l'image courante, $(x^{'}, y^{'})$ et (p_1, p_3, p_4) les 3 paramètres du modèle.
Ce modèle permet de caractériser des mouvements moins complexes de la caméra uniquement dans l'espace 2D, les 3 mouvements suivant l'axe z n'étant pas modélisés. Dans un premier temps les auteurs ont utilisé le modèle à 3 paramètres pour estimer le mouvement de la caméra en utilisant l'algorithme d'estimation rapide présenté dans [TSKR00]. Après quoi, l'article démontre que pour obtenir une bonne modélisation du mouvement de la caméra il convient mieux d'utiliser le modèle à 6 paramètres. C'est pourquoi, les auteurs proposent une détection des frontières de changement des plans de montage utilisant à la fois les valeurs des 6 paramètres de mouvement estimés, le calcul de l'erreur moyenne quadratique des vecteurs de mouvement conformes au modèle et du nombre de macro-blocs codés en intra dans les images de type P. L'algorithme présenté pour la détection des transitions vidéo commence par calculer la différence en valeur absolue entre les 6 paramètres du mouvement estimés dans les images de type P :

$$\Delta a_n(k) = |a_n(k) - a_n(k-1)|,\ n = 1, \ldots, 6 \tag{33}$$

avec $a_n(k)$ le paramètre de mouvement de l'image k et d'indice n.

Puis, les valeurs de Δa_n sont normalisées par rapport à une valeur de référence Δ_{ref} qui peut être soit la moyenne des valeurs, soit le maximum des valeurs. Après quoi, les auteurs définissent une seconde fonction, $A(k)$, définie comme suit :

$$A(k) = \sum_{n=1}^{6} \Delta^* a_n(k) \tag{34}$$

avec $\Delta^* a_n$ la valeur normalisée de Δa_n.
De plus, les auteurs considèrent les deux fonctions suivantes :

- $Q(k)$ qui correspond au nombre de macro-blocs codés en intra dans l'image de type P numéro k et
- $MSE(k)$ qui représente la somme des erreurs quadratiques dans l'image de type P numéro k.

Enfin, l'ensemble des trois fonctions définies plus haut sont regroupées en une seule sous forme d'une combinaison linéaire. La nouvelle fonction, $D(k)$ ainsi formée est définie par :

$$D(k) = (Q(k) + \alpha)(MSE(k) + \alpha)(A(k) + \alpha) \tag{35}$$

avec $\alpha > 0$ un paramètre le plus souvent fixé à 1.
C'est précisément l'évolution des valeurs de D que les auteurs étudient afin d'en déduire la présence ou non d'une frontière de plan de montage. Pour éviter un nombre trop important de fausses alarmes les auteurs supposent notamment qu'un plan doit avoir une durée au moins égale à 5 images, $l_{min} = 5$, de type P, ce qui correspond à 0.5 seconde au format PAL/SECAM. Il est aussi important de vérifier que le rapport entre deux valeurs $D(i-1)$ et $D(i)$ correspondant à deux pics successifs est supérieur à un seuil fixé. Ainsi, si le rapport $D(i)/D(i-1)$ est supérieur au seuil alors $D(i)$ est validée comme étant une rupture dans le contenu vidéo.

Les tests menés dans l'article ont été effectués sur des extraits très courts variant entre 13 et 36 secondes par vidéo. La base de données de test comprenait 3 vidéos. Les meilleures performances de la méthode sont de l'ordre de 94% de rappel et 89% de précision. Notons, tout de même, que ces valeurs ont été obtenues à partir d'une base de données vidéo trop petite pour avoir une idée réelle des performances de la méthode. Toutefois, ceci vient du fait que le temps de calcul de l'estimation de mouvement est très lourd. C'est la raison pour laquelle les auteurs n'ont pas évalué leur méthode sur une plus grande base.

Primaux, Benois-Pineau, Krämer et Domenger [PBPKD04]

Dans [PBPKD04], les auteurs proposent une méthode pour la détection des frontières de plan de montage basée sur le paradigme du *rough indexing*. Ce paradigme consiste à extraire un minimum d'information du flux vidéo compressé. Il s'agit d'une part de proposer un algorithme de traitement temps réel et d'autre part de réduire au maximum la dimension de l'espace des descripteurs. La méthode proposée permet de détecter les frontières des plans pour les images de type I et P. La méthode de détection des frontières sur les images P est une amélioration de la méthode proposée dans [DBP01]. Les résultats expérimentaux obtenus sont assez bons comparés aux résultats obtenus par les autres participants à la campagne TRECVID 2004 et

sachant que les données d'entrée représentent les vecteurs de mouvement MPEG qui peuvent être loin de la réalité du mouvement dans la séquence. La meilleure précision est de 73% pour un rappel de 65% et le meilleur rappel est de 74% pour une précision de 57%.

2.1.3 Comparatif des différentes méthodes

La comparaison des diverses approches pour la détection de changement des plans n'a de sens que si ces détecteurs ont été testés sur un grand corpus de données suffisamment riche aussi bien en quantité, type de transitions, qu'en qualité et résolution de la vidéo. C'est le cadre de la campagne TRECVID. Néanmoins, il est difficile de privilégier tel ou tel groupe de méthodes selon leurs résultats dans TREC. En réalité, les résultats les plus probants sont obtenus par la fusion des deux analyses : spatiale et temporelle. Il est évident que les détecteurs de changements de plans doivent être accompagnés par la détection des effets spéciaux (flash d'appareil photo) et que les algorithmes ad-hoc doivent être proposés pour chaque type de transition graduelle tenant compte de son modèle présumé (volet, fondus enchaînés, etc...). Il nous semble intéressant de citer ici l'approche plus globale qui, comme le détecteur basé sur le paradigme de *rough indexing* [NEP+05] s'intéresse plutôt à la détection de toutes les transitions sans distinction.

Dans [Han02], l'auteur propose de développer un détecteur de transitions vidéo à la fois générique, sans a priori sur le support vidéo, et le plus robuste possible. Il s'agit de combiner en un seul et même système l'ensemble des méthodes déjà connues afin de tirer profit des avantages de chacune.
L'auteur rappelle plus en détail les différentes grandes méthodes développées auparavant. Il y a principalement deux grandes familles d'algorithmes. Les approches basées sur l'extraction et l'analyse de descripteurs visuels tels que la valeur de luminance, de chrominance, le calcul des histogrammes de couleur, etc... et les approches utilisant la compensation du mouvement.

Les méthodes les plus simples pour évaluer et quantifier les discontinuités dans les flux vidéo synonymes de la présence de transitions vidéo sont les méthodes basées sur l'extraction et l'analyse des descripteurs vidéo et comparaison entre deux images consécutives des variations de ces valeurs. L'auteur prend l'exemple de la moyenne des valeurs absolues des différences des intensités de chaque pixel pour tous les couples d'images successives de la vidéo. Ces valeurs sont comparées à un seuil pour décision de la présence ou non de transitions vidéos. Cette règle de décision peut-être écrite sous la forme suivante :

$$z(k,k+1) = \frac{1}{XY}\sum_{x=1}^{X}\sum_{y=1}^{Y} D_{k,k+1}(x,y) \tag{36}$$

avec $D_{k,k+1}(x,y) = \begin{cases} 1, \text{ si } |I_k(x,y) - I_{k+1}(x,y)| > T_1 \\ 0, \text{ sinon} \end{cases}$ et $I(x,y)$ l'intensité lumineuse du point de coordonnées (x,y), T_1 un seuil prédéfini et X et Y les dimensions totales de l'image considérée.

Les méthodes qui se sont développées par la suite sont celles basées sur la compensation de mouvement pour tous les couples d'images de la vidéo. Pour obtenir cette compensation de mouvement il faut appliquer une procédure de mise en correspondance des blocs d'une image d'indice k avec une autre image d'indice $k+l$. La méthode de mise en correspondance des

blocs peut être formalisée par l'équation suivante :

$$\begin{aligned} D_{k,k+l}(i) &= D(b_i(k), b_{i,m}(k+l)) \\ &= min_{j=1...N} D(b_i(k), b_{i,j}(k+l)) \end{aligned} \tag{37}$$

avec D une valeur de différence absolue, $b_i(k)$ le bloc d'indice i de l'image d'indice k, N le nombre total de blocs de l'image.
Lorsque deux images appartiennent au même plan alors la valeur $D_{k,k+l}(i)$ sera plutôt faible, ce que ne sera pas le cas si une transition vidéo intervient entre les deux images d'indice k et $k+l$. Ainsi, les variations des valeurs de $D_{k,k+l}(i)$ sont de bons indices de la présence de transitions vidéo. Il est aussi possible de calculer d'autres valeurs de cohérence basées sur D. Notamment, la méthode qui consiste à normaliser les valeurs de D obtenues. Ces nouvelles valeurs sont notées $d^s_{k,k+l}(i)$. Puis on applique un coefficient pondérateur c_i sur ces mêmes valeurs. Enfin, la nouvelle valeur de discontinuité, $z(k, k+l)$, ainsi obtenue peut être formalisée de la manière suivante :

$$z(k, k+l) = \sum_{i=1}^{N} c_i d^s_{k,k+l}(i) \tag{38}$$

avec N le nombre total de blocs dans l'image.

Après avoir adressé le problème de la détection des frontières des plans de montage puis présenté les différentes méthodes existantes pour résoudre ce problème, l'auteur propose une nouvelle approche en tenant compte des méthodes déjà existantes.

Ainsi, la méthode proposée par l'auteur est basée sur un modèle statistique. Le principe utilisé est celui du calcul du maximum de vraisemblance pour deux hypothèses données afin de décider laquelle des deux hypothèses est la plus probable. Les deux hypothèses considérées ici sont :

- Hypothèse S : Présence d'une transition entre les images k et $k+l$
- Hypothèse $\bar{S}$: Absence d'une transition entre les images k et $k+l$

L'auteur définit ensuite les probabilités de manquer une transition et de fausses détections comme :

$$P_M = \int_{Z_{\bar{S}}} p(z|S)dz \tag{39}$$

$$P_F = \int_{Z_S} p(z|\bar{S})dz \tag{40}$$

avec $Z_{\bar{S}}$ et Z_S les intervalles des valeurs de discontinuité relativement aux hypothèses S et $\bar{S}$, $p(z|S)$ et $p(z|\bar{S})$ les fonctions de vraisemblance associées relativement aux hypothèses S et $\bar{S}$. L'auteur propose de formuler la probabilité d'erreur moyenne de détection comme suit :

$$\begin{aligned} P_E &= P_E(k) \\ &= P_k(S)P_M + P_k(\bar{S})P_F \\ &= P_k(S) \int_{Z_{\bar{S}}} p(z|S)dz + P_k(\bar{S}) \int_{Z_S} p(z|\bar{S})dz \end{aligned} \tag{41}$$

avec $P_k(S)$ et $P_k(\bar{S})$ les probabilités pour les hypothèses S et $\bar{S}$ pour la paire d'images d'indice k et $k+l$.

Comme les hypothèses S et $\bar{S}$ forment une partition de l'ensemble des hypothèses alors nous avons :

$$P_k(\bar{S}) = 1 - P_k(S) \tag{42}$$

Puis, l'auteur définit la probabilité $P_k(S)$ comme le produit de la probabilité *a priori*, $P_k^a(S)$, pour l'hypothèse S et de la probabilité conditionnelle, $P_k(S|\psi(k))$, pour la même hypothèse entre les images k et $k+l$. D'où la formulation suivante :

$$Pk(S) = P_k^a(S)P_k(S|\psi(k)) \tag{43}$$

avec $\psi(k)$ des informations supplémentaires extraites de la vidéo.
Ainsi pour maximiser la robustesse du détecteur, il faut minimiser la probabilité d'erreur moyenne P_E. Pour cela, il faut calculer le rapport entre les fonctions de vraisemblance associées à chacune des hypothèses. Ce rapport est alors comparé à un seuil probabiliste afin de décider laquelle des deux hypothèses est la plus probable dans la situation donnée. La règle de décision finale peut être formulée de la manière suivante :

$$\frac{p(z|S)}{p(z|\bar{S})} \begin{array}{c} \bar{S} \\ < \\ > \\ S \end{array} \frac{1 - P_k(S)}{P_k(S)} = \frac{1 - P_k^a(S)P_k(S|\psi(k))}{P_k^a(S)P_k(S|\psi(k))} \tag{44}$$

avec $P_k^a(S)$ et $P_k(S|\psi(k))$ estimées à partir de l'équation (43).

Les phases de test et de validation du détecteur proposé ont été réalisées sur une base de données vidéo comprenant 5 séquences de 4 genres vidéo différents : film, match de football, journal télévisé et documentaire. L'ensemble de ces séquences comprend 104 transitions abruptes et 23 graduelles. Les performances globales affichées par le détecteur proposé sont résumées dans le tableau de la figure 2. Ces résultats sont très satisfaisants, toutefois ils ont

	Rappel	Précision
Transitions abruptes	100%	100%
Transitions graduelles	83%	79%

FIG. 2 – Performances globales du détecteur statistique

été obtenus sur une base de données de petite taille. Il faudrait valider ces résultats à plus grande échelle, sur les bases de données de TRECVID par exemple.

2.2 Segmentation en scènes

Après avoir présenté le découpage des contenus multimédia en plans de montage, nous allons maintenant adresser le problème de la segmentation en scènes basée sur l'analyse uniquement du flux vidéo. Deux approches principales sont principalement utilisées dans la littérature :

- l'approche selon laquelle une scène est modélisée par un groupement des plans de montage dont le contenu est considéré comme similaire [TZ04, VRB00, YL95, YYL96].
- l'approche selon laquelle une scène est délimitée par deux frontières successives [NLBP+04, KGOG03, KGG+03].

Ainsi, de nombreux auteurs s'attachent à proposer des modèles spécifiques de scène comme, entre autres, une détection des scènes de dialogue [AAW01], des scènes d'intérieur et d'extérieur [MHI+02]. Les scènes sont d'abord détectées puis classifiées. Toutes ces approches sont supervisées, elles reposent sur un apprentissage statistique au préalable. Des travaux ont aussi été menés quant à la segmentation en scènes génériques de contenus multimédia spécifiques tels que les journaux télévisés [HClK+03] dans le cas de la campagne d'évaluation TRECVID, ou bien les films [RS03], ou encore les documents sportifs [LMP04, HS03, KGOG03, KGG+03]. Proposer des méthodes basées sur un modèle de scène générique sans restriction sur le support audiovisuel est un réel challenge scientifique [AAD+03, CSLZ02, LWC98].

Miene, Hermes, Ioannidis, Fathi et Herzog [MHI+02]

La méthode proposée dans [MHI+02] consiste à détecter les frontières des plans de montage puis à classifier ces plans comme des scènes d'intérieur ou d'extérieur. L'ensemble de ces travaux ont été réalisés dans le cadre de la campagne d'évaluation TREC 2002. Tout d'abord, la méthode propose un algorithme de détection des plans basé, dans un premier temps, sur l'extraction de descripteurs vidéo puis, dans un second temps, sur la détection des plans au regard des descripteurs extraits précédemment. L'extraction des descripteurs vidéo commence par une conversion des images couleurs en images en niveaux de gris puis l'histogramme, H_G, est calculé ainsi que l'histogramme, $H_{G_{Diff}}$, qui correspond à la différence quadratique des histogrammes de deux images consécutives. Cette mesure est définie par :

$$H_{G_{Diff}}(n, n-1) = \sum_{i=0}^{255} \frac{(H_G(n)(i) - H_G(n-1)(i))^2}{Max(H_G(n)(i), H_G(n-1)(i))} \tag{45}$$

avec $H_G(n)(i)$ la valeur de l'histogramme en niveaux de gris à l'index i de l'image n, $Max(H_G(n)(i), H_G(n-1)(i))$ le maximum entre les deux valeurs des histogrammes $H_G(n)(i)$ et $H_G(n-1)(i)$.
Ce maximum est aussi utilisé comme vecteur de normalisation. Les valeurs de $H_{G_{Diff}}$ obtenues sont alors comparées à un seuil adaptatif défini par :

$$Th = \frac{Max\{H_{G_{Diff}}(1,0), \ldots, H_{G_{Diff}}(n, n-1)\}}{100} \tag{46}$$

Dans le cas des transitions graduelles, la méthode génère plusieurs fausses alarmes pour un même plan mais les auteurs apportent une solution qui consiste à considérer ces fausses alarmes comme appartenant au même plan. Pour cela, la méthode proposée calcule une distance temporelle entre ces fausses alarmes et regroupe celles dont la distance temporelle est inférieure à un seuil. L'image d'index le plus faible est considérée comme l'image de début et l'image d'index le plus fort comme l'image de fin de la transition graduelle ainsi reconstituée.

L'étape suivante consiste à classifier les plans de montage précédemment détectés en scènes d'intérieur ou d'extérieur. Les descripteurs utilisés pour cette classification sont la moyenne, la variance et le nombre de pics pour chaque histogramme : l'histogramme des 3 composantes couleur (RGB) et l'histogramme du niveaux de gris. Les auteurs utilisent un réseau de neurones qui a été, dans un premier temps, entraîné avec les descripteurs statistiques décrits ci-dessus. Pour pouvoir classifier les plans à partir des descripteurs, la méthode consiste à extraire n images clés dans un même plan vidéo i puis à les soumettre au réseau de neurones à deux

sorties, une sortie par classe. Toutefois, les auteurs ne considèrent pas les images au voisinage des transitions afin d'éviter d'éventuelles erreurs de classification dans le cas de transitions graduelles. À la sortie du réseau de neurones, deux listes de n valeurs correspondant à chacune des sorties du réseau sont obtenues. La valeur médiane est calculée pour chacune des listes afin d'obtenir, pour le plan i, les probabilités d'appartenance à la classe Intérieur ou Extérieur. Enfin, la décision finale est prise à partir du calcul de la différence entre les valeurs médianes des deux listes. Si cette différence est supérieure à un seuil, alors le plan est classifié suivant la classe dont la probabilité d'appartenance est la plus grande.

Les résultats présentés dans le papier ne font référence qu'au détecteur de frontières des plans de montage, la classification en scènes d'intérieur et d'extérieur n'étant pas représentée. Les expérimentations pour l'évaluation des performances du détecteur de transitions vidéo ont été menées sur un corpus vidéo comprenant 344 transitions vidéo. Les résultats obtenus sont de l'ordre de 66% de rappel pour 94% de précision. Ces résultats incluent la détection des transitions abruptes et graduelles sans distinction. Le faible taux de rappel s'explique par les erreurs de détection des transitions graduelles.

Rasheed et Sahah [RS03]

Dans [RS03] , les auteurs proposent une approche pour la détection des scènes dans les films hollywoodiens et les spectacles de variétés. Le schéma global de la méthode est exposé par la figure suivante : Comme le montre la figure 3, le principe général de la méthode consiste à rechercher les frontières des plans de montage puis à extraire des descripteurs bas-niveau pour ces mêmes plans. À partir de ces descripteurs, les auteurs proposent une approche en deux passes pour obtenir les frontières de scène. La première de ces passes consiste à calculer une mesure de similarité de couleur entre les plans (BSC). Les fortes variations de ces valeurs sont considérées comme des possibles frontières de scène (PSB). C'est lors de la seconde passe que les possibles frontières de scène détectées précédemment sont validées ou rejetées par l'estimation de la dynamique des scènes détectées lors de la première passe. La dynamique de scène est modélisée comme une combinaison linéaire entre la longueur de la scène et l'estimation du mouvement à partir des plans de la scène. Toutes ces notions sont développées par la suite. La méthode utilisée pour la détection des frontières des plans de montage est basée sur la calcul de l'intersection des histogrammes couleur entre deux images successives de la vidéo. Cette mesure est définie par :

$$D(f^i, f^j) = \sum_{b \in bins} min(H_i(b), H_j(b)) \tag{47}$$

avec $f^{i,j}$ les images d'index i et j et $H_{i,j}$ les histogrammes correspondant aux images d'index i et j.
De manière simple, une frontière de plan est détectée lorsque :

$$D(f^i, f^{i-1}) < T_{color} \tag{48}$$

avec T_{color} un seuil fixe.
Après avoir caractérisé les frontières des plans de montage, les auteurs proposent un algorithme d'extraction des images clés pour chacun des plans. Cet algorithme considère dans un premier temps l'image centrale du plan comme la première image clé, K_1. Les images clés suivantes sont extraites selon la règle suivante : chaque image du plan est comparée à chaque image clé

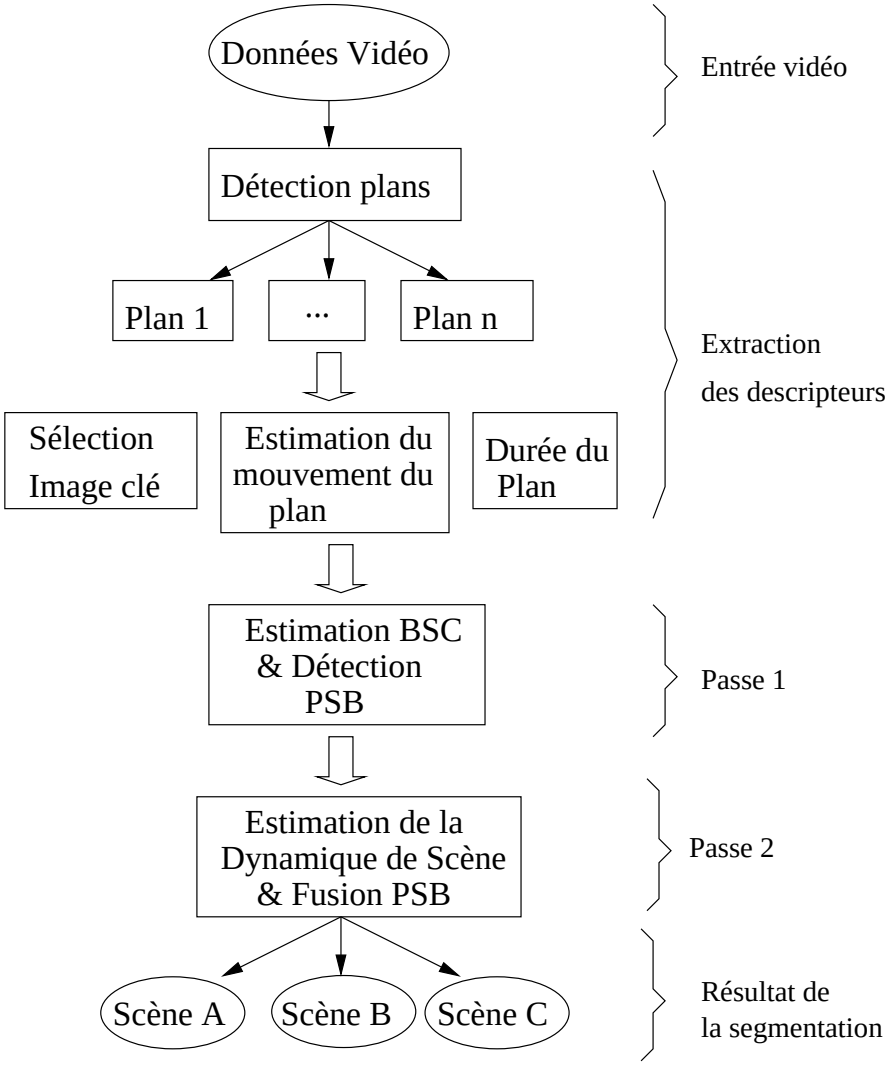


Fig. 3 – Schéma général de la méthode de détection des frontières de scènes : BSC = Backward Shot Coherence et PSB = Potential Scene Boundaries

déjà détectée. Si l'image diffère de trop par rapport aux images clés précédemment détectées alors elle est ajoutée à l'ensemble des images clés sinon elle est ignorée. Cet algorithme peut s'écrire sous la forme :

$$\forall j \in \{a, a+1, \ldots, b\}, max(D(f^j, f^k)) < Th \quad \forall f^k \in K_i \tag{49}$$

avec a et b respectivement l'indice des images de début et de fin d'un plan, K_i la i-ème image clé de l'ensemble des images clés et Th un seuil minimal fixé.

L'étape suivante (voir figure 3) consiste en l'extraction des descripteurs vidéo. Il faut tout d'abord estimer le mouvement global de la caméra, pour cela les auteurs utilisent le modèle affine du mouvement à 6 paramètres déjà présenté (voir équation (31)). Puis ils estiment

l'amplitude de l'erreur d'estimation grâce à l'équation suivante :

$$\epsilon_j = \sum_{k \in blocs} \sqrt{(u_k^{'} - u_k)^2 + (v_k^{'} - v_k)^2} \tag{50}$$

avec u_k, v_k les coordonnées du vecteur vitesse encodé et $u_k^{'}$, $v_k^{'}$ les dérivées des coordonnées du vecteur vitesse du $k_{ième}$ bloc dans l'image d'index j.
Enfin, les auteurs définissent un descripteur vidéo, qu'ils appellent le contenu du mouvement d'un plan, défini sur les images P par :

$$SMC_i = \sum_{j \in S_i} \epsilon_j \tag{51}$$

La dernière phase de la méthode est constituée de l'algorithme à deux passes de détection des scènes. La première passe calcule une mesure de cohérence entre deux plans, i et j, définie par :

$$SC_i^j = max_{f^x \in K_i, f^y \in K_j}(D(f^x, f^y)) \tag{52}$$

avec $K_{i,j}$ l'ensemble des images clés des plans i et j.
Une seconde mesure découle directement de la première, il s'agit de la cohérence de plan vers l'arrière. Cette nouvelle valeur correspond au maximum des valeurs de SC (voir équation 52) comprises dans une fenêtre de taille N. D'où :

$$BSC_i = max_{1 \geq k \geq N}(SC_i^{i-k}) \tag{53}$$

avec BSC_i la cohérence du plan vers l'arrière pour le plan i.

À partir de l'analyse des valeurs de BSC (voir équation 53), les auteurs ont remarqué qu'elles étaient assez stables au cours d'une scène mais qu'au contraire elles subissaient de fortes variations dans le cas d'un changement de scène. Ainsi, les auteurs localisent les maxima locaux et les étiquettent comme étant une frontière de scène potentielle (PSB). Dans certains cas, cette méthode génère des fausses alarmes notamment dans le cas où de nouvelles personnes apparaissent (ou disparaissent) au sein d'un même plan. Pour filtrer ces petites erreurs, les auteurs proposent le test suivant :

$$D(f^i, f^j) \geq T_{color} \tag{54}$$

avec $f^i \in Scène_k$ et $f^j \in Scène_{k+1}$.
Si l'inéquation 54 est vérifiée, alors la frontière de scène détectée est considérée comme une fausse alarme et elle est alors annulée.

Enfin, lors de la seconde passe, les auteurs proposent une nouvelle mesure afin d'éliminer un nombre plus important de fausses détections. En effet, les scènes dont l'activité de mouvement est assez calme sont assez bien caractérisées lors de la première passe. En revanche, les scènes dont le contenu est plus chaotique et qui contiennent des parties avec des zones de fort mouvement contiennent des fausses alarmes qui n'ont pas été éliminées lors de la première passe. Pour cela, les auteurs calculent la dynamique de scène pour chacune des scènes potentielles détectées lors de la première phase. Ce descripteur vidéo est calculé à partir de la formule suivante :

$$SD_i = \frac{\sum_{j \in Scène_i} SMC_j}{\sum_{j \in Scène_i} L_j} \tag{55}$$

avec SD_i la valeur de la dynamique de scène i, SMC_j le contenu du mouvement du plan j dans la même scène et L_j la longueur du plan correspondant.
Dans le cas où la valeur de SMC est grande pour une faible valeur de L une valeur haute est obtenue pour SD, ce qui signifie une forte dynamique dans la scène. Ainsi, la mesure de la dynamique de scène est calculée pour chacune des scènes potentielles. Puis si deux valeurs de SD relatives à deux scènes potentielles adjacentes sont supérieures à un seuil fixé alors la frontière entre ces deux scènes est éliminée.

Les résultats ont été obtenus sur des extraits de 5 films hollywoodiens dont la durée varie entre 35 et 60 minutes. Ces extraits ont été pris à partir du milieu de film global. D'autres extraits ont été utilisés et sont issus de séries télévisées et d'une émission de variétés. Les performances affichées sont très satisfaisantes puisqu'elles varient de 86 à 88% de rappel et de 63 à 81% pour la précision. Ces résultats sont très bons compte tenu de la durée et de la variété des contenus testés. Toutefois, les auteurs se sont restreints à des genres vidéos particuliers et n'ont pas tenté de proposer une généralisation de leur méthode à d'autres contenus comme les contenus sportifs ou bien les clips musicaux.

Nicolas, Manoury, Benois-Pineau, Dupuy et Barba [NMBP+04]

Dans [NMBP+04], les auteurs proposent une méthode de détection des scènes par regroupement des plans de montage. La méthode proposée est basée sur l'utilisation des mosaïques 1D comme descripteur vidéo. Le principe consiste à signer chaque plan de montage détecté en utilisant le descripteur précédemment cité, ainsi il est possible d'établir une distance entre deux plans. Cette distance est synonyme de similarité, plus la distance est faible, plus la similarité des deux plans considérés est grande.
Cette méthode s'applique aux contenus de type sportif ou bien journaux télévisés. Pour ces types de contenus il est très difficile de proposer un modèle de scènes. Les résultats expérimentaux obtenus par cette méthode sont prometteurs et demandent à être validés sur un corpus vidéo de test plus riche.

2.3 Conclusion

Dans ce chapitre, nous avons étudié diverses méthodes d'indexation des documents vidéo à l'execption des méthodes d'analyse cross-média présentées dans le chapitre suivant. Partant du problème fondamental de détection robuste de changement des plans, les méthodes d'indexation visent à remonter à un autre niveau sémantique, celui des scènes. En récapitulant l'analyse bibliographique de ces méthodes nous pouvons dégager deux grandes tendances :

- les méthodes par regroupement à partir des unités de base (plans de montage) et
- les méthodes identifiant parmi toutes les frontières des segments vidéo, celles qui correspondent aux changements de scènes. Cette approche peut être qualifiée de différentielle.

Enfin, il paraît pertinant de fusionner les informations provenant des flux audio et vidéo afin de proposer une segmentation en scènes plus robuste que les approches monomédia. Dans le chapitre 4, nous nous intéressons donc aux travaux portant sur l'analyse cross-média audio et vidéo d'un document multimédia.

Chapitre 3

État de l'art de l'analyse et de l'indexation des flux audio

Dans cette partie, nous développons et étudions les différentes méthodes présentes dans la littérature dans le domaine de l'analyse et de l'indexation des flux audio. L'analyse audio se décompose en deux niveaux :

- **les descripteurs de bas niveau :** partiels, formants, harmonicité, enveloppe spectrale, etc...
- **les descripteurs de haut niveau :** style musical, rythme, orchestration, mélodie, harmonie, etc...

Les descripteurs de bas niveau servent à modéliser les instruments dans le cas de l'analyse dans le domaine musical mais de manière générale ils permettent d'extraire les caractéristiques d'un signal audio. À partir de ces descripteurs, il est possible de mettre au point des méthodes comme par exemple :

- reconnaître Musique/Non-Musique, Parole/Non-Parole
- classifier une bande sonore en Parole/Musique/Bruit/Silence

En ce qui concerne les descripteurs de haut niveau, ils permettent de modéliser le morceau de musique joué par un instrument ou plus généralement de caractériser un contenu audio par des informations de haut niveau sémantique : nom de l'auteur pour un morceau de musique ou bien le nom du locuteur pour un segment de parole, etc...
Plus concrètement, les fichiers de données sur lesquels nous avons travaillé sont des fichiers au format WAV. Ce type de fichiers qui contiennent les échantillons de l'onde sonore donc une analyse au niveau des descripteurs de bas niveau s'impose, contrairement aux fichiers MIDI qui contiennent déjà les notes (volume, hauteur, durée) pour lesquels une analyse des descripteurs de bas niveau n'est pas nécessaire.

Dans la section 3.1, nous présentons les principaux descripteurs bas niveau utilisés pour l'indexation audio. Dans la section suivante (section 3.2), nous étudions les principales méthodes utilisées pour l'indexation des flux audio.

3.1 Analyse des flux audio numériques

Dans cette section, nous étudions les différents descripteurs audio couramment utilisés dans la littérature. Nous présentons ainsi l'intérêt de chacun en fonction du problème à résoudre.

La première sous-partie de cette section est consacrée à l'introduction du domaine de l'analyse des flux audio numériques. Dans la seconde sous-section, nous présentons quelques éléments de traitement du signal numérique. C'est précisément dans cette sous-section que nous présentons les différents descripteurs audio les plus couramment utilisés dans la littérature. Enfin, les sous-sections suivantes sont consacrées à la présentation des travaux existant dans les domaines de la parole, de la musique, du silence et du bruit.

3.1.1 Introduction

Le domaine sonore est composé de quatre grandes classes :

- la parole,
- la musique,
- le silence et
- le bruit qui contient tout ce qui n'est ni de la parole, ni de la musique, ni du silence.

Les deux premières classes ont fait l'objet du plus grand nombre de travaux [TC99, MB03, ZK01, ZP04] à l'heure actuelle. Les travaux portant sur l'analyse des silences ont présenté un faible intérêt [JEA99, Sou83] du fait de la simplicité de cette classe et du nombre restreint d'applications. Enfin, encore peu de travaux dans la littérature [BPJ02] traitent de l'analyse du bruit. Ceci étant dû, en majorité, à l'aspect complexe d'une telle classe car il est difficile de la modéliser du fait de sa grande richesse et diversité.

3.1.2 Éléments de traitement du signal numérique

Un son pur est un signal de forme sinusoïdale, de fréquence et d'amplitude constante, ces sons n'existent pas à l'état naturel, on ne peut les obtenir qu'avec des générateurs électroniques. En réalité, les sons qui arrivent à nos oreilles sont composés de plusieurs sinusoïdes de fréquences et d'amplitudes différentes. Ces fréquences sont appelées *harmoniques* si elles sont multiples de celles que nous entendons qui est l'harmonique 1 ou *fondamentale*. Les autres étant les *partiels*. Un exemple connu riche en harmoniques et en partiels est celui d'une cloche. Ces harmoniques peuvent ne pas démarrer aux mêmes moments, on parlera alors de *déphasage*. Dans le cas de la parole, nous pouvons voir la présence de *phonèmes* au niveau du signal sonore. Un phonème est la plus petite entité du langage parlé.

Un son peut-être représenté sous deux formes : temporelle et fréquentielle. Dans le cas d'une représentation temporelle, le signal est modélisé comme des variations continues de l'amplitude en fonction du temps, t, mesurées en secondes. La représentation fréquentielle est obtenue grâce à la transformée de Fourier. En effet, il est possible de décomposer un signal complexe variant en amplitude en fonction du temps en une somme de sinusoïdes ayant chacune une fréquence et une amplitude variant dans le temps (figure 4). Le signal obtenu par la transformée est aussi appelé le *spectre*. Le spectre permet de déterminer si le signal provient d'une source sonore *harmonique*, généralement musicale mais aussi vocale suivant les lettres

prononcées, ou d'une source *in-harmonique*. L'ensemble des éléments du spectre d'un signal est appelé l'enveloppe spectrale (voir figure 5).

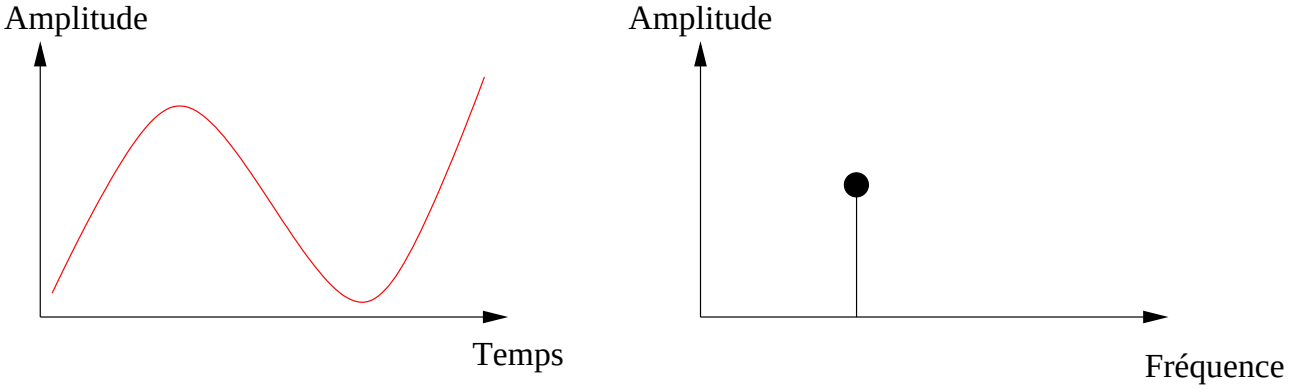


Fig. 4 – Deux représentations : temporelle et fréquentielle.

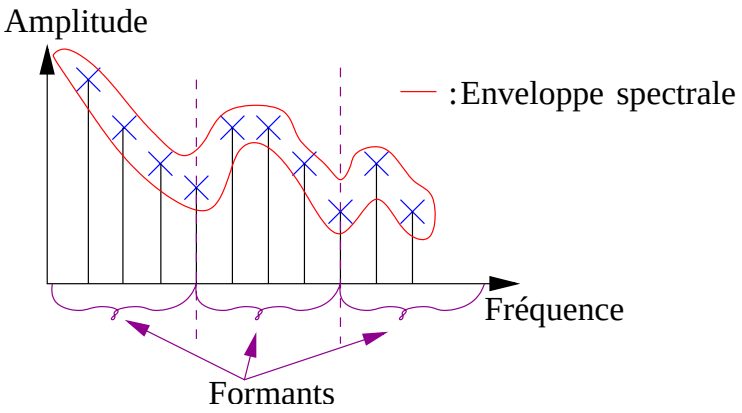


Fig. 5 – Exemple d'enveloppe spectrale

Afin de pouvoir être traité de manière numérique, un signal analogique doit être numérisé. Le principe de la conversion est basée sur une quantification des dépressions d'air que crée l'onde sonore à des intervalles de temps réguliers. Ces intervalles définissent la fréquence d'échantillonnage. Le signal obtenu est un signal discret par opposition au signal continu dans le cas d'un signal analogique.
Un point crucial dans le processus de numérisation d'un signal est l'échantillonnage. En effet, pour que le signal puisse conserver ses propriétés d'origine il y a des règles à respecter. L'une des règles de base est la nécessité d'avoir au moins deux échantillons par période du signal pour avoir les informations suffisantes concernant les fréquences qui le composent. C'est la raison pour laquelle il faut s'assurer que la fréquence d'échantillonnage soit au moins le double de la fréquence du signal échantillonné. La fréquence limite est appelée *fréquence de Nyquist*[Roa98]. Comme l'oreille humaine n'est pas sensible aux fréquences supérieures à 20kHz, la fréquence d'échantillonnage doit être supérieure à 40kHz.

L'analyse audio nécessite des notions de traitement du signal comme la transformée de Fourier. Une présentation de ces notions fondamentales est donnée en annexe (voir Annexe A).

3.1.3 Descripteurs statistiques du signal pour l'analyse audio

Dans cette partie, nous nous intéressons aux descripteurs statistiques du signal pour la modélisation des signaux numériques. Nous présentons les principaux concepts les plus répandus dans la littérature ainsi que ceux utilisés dans la suite de ce document.

Variance - Écart-type

La variance est un moment statistique d'ordre 2. La variance permet de caractériser la dispersion des valeurs d'un signal autour de la moyenne. L'écart-type est obtenu en prenant la racine carrée de la variance. D'où :

$$\sigma = \sqrt{V(X)} = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(X[i]-\mu)^2} \tag{56}$$

avec V la variance, σ l'écart-type, N le nombre total d'échantillons du signal X dont la moyenne vaut μ.

Skewness

Cette mesure, moment statistique d'ordre 3, est définie par :

$$S_X = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(X[i]-\mu)^3 \tag{57}$$

avec S_X la valeur du skewness pour un signal X contenant N échantillons dont la moyenne est μ et l'écart-type σ.
Cette grandeur caractérise la symétrie de la densité de probabilité. Une variable aléatoire de densité de probabilité symétrique à un skewness nul. Plus la densité de probabilité est asymétrique, plus le skewness admet une grande valeur. Quelques valeurs pour des distributions classiques sont données par le tableau de la figure 6.

Kurtosis

Le kurtosis est un moment statistique d'ordre 4 défini par :

$$K_X = \frac{1}{N\sigma^4}\sum_{i=1}^{N}(X[i]-\mu)^4 \tag{58}$$

avec K_X la valeur du kurtosis pour un signal X contenant N échantillons dont la moyenne est μ et l'écart-type σ.

Le kurtosis permet de caractériser la forme de la densité de probabilité. Plus le kurtosis est important, plus la forme de la fonction de densité de probabilité est étroite. Au contraire, plus la valeur du kurtosis est faible plus la forme de la fonction de densité de probabilité est plate. Quelques valeurs pour des distributions classiques sont données par le tableau de la figure 6.

Distribution	Skewness	Kurtosis
Uniforme	0.0	1.8
Gaussienne	0.0	3.0
Rayleigh	0.6	3.2

Fig. 6 – Exemple de valeurs de skewness et kurtosis

Le Centroïde Spectral

Cette grandeur caractérise le barycentre du spectre d'un signal. Il est calculé de la manière suivante :

$$\frac{\sum_{i=0}^{\frac{N}{2}} X[i] \times i}{\sum_{i=0}^{\frac{N}{2}} X[i]} \tag{59}$$

avec $X[i]$ le i-ème élément du spectre X d'un signal contenant N échantillons.

Le centroïde spectral est très utilisé dans le domaine de la parole, il permet de caractériser les parties voisées des parties non-voisées. Il permet ainsi de séparer les parties voisées du bruit [PRMAO03].

MFCC

Les coefficients cepstraux de l'échelle Mel (voir section 3.1.5), appelés MFCC pour Mel-Frequency Cepstral Coefficient. Ils ont été développés par Rabiner [RJ93] pour modéliser la parole. Généralement, seuls les treize premiers coefficients cepstraux sont considérés. D'autant plus que le premier coefficient correspond à l'énergie du signal (voir section 3.1.5) à un facteur près. Cette grandeur caractérise la forme du spectre obtenu par la transformée de Fourier. Le diagramme de la figure 7 illustre la manière dont sont calculés les MFCC :

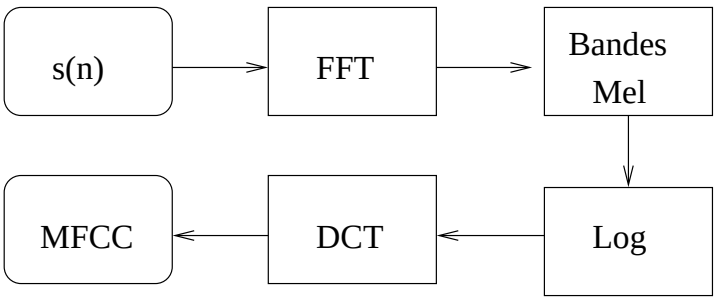


Fig. 7 – Schéma de calcul des MFCC

Le Roll-Off Spectral

Le Roll-Off spectral ou point de roulement du spectre correspond à la fréquence pour laquelle 95% de la distribution en puissance spectrale est inférieure à cette même fréquence. Ce descripteur est utilisé, principalement, pour distinguer les parties voisées des parties non

voisées de la parole et de la musique. Le roll-off spectral est défini par :

$$\sum_{i=0}^{f_c} X[i] = 0.95 \sum_{i=0}^{sr/2} X[i] \tag{60}$$

avec f_c la fréquence du point de roulement du spectre, $X[]$ le spectre du signal et $sr/2$ la fréquence de Nyquist [Roa98].

Le taux de passage par zéro

Le taux de passage par zéro est le nombre de fois où le signal, dans sa représentation temporelle, coupe l'axe des abscisses (axe temporel). Cette mesure est corrélée au centroïde spectral, il est donc aussi utilisé dans les travaux de segmentation en parole/musique [LJZ01]. Par exemple dans le cas de la parole, les parties voisées ont plutôt tendance à avoir un taux de passage par zéro faible (voir figure 8), contrairement aux parties non voisées qui affichent une valeur élevée (voir figure 9). Les deux figures 8 et 9 illustrent les valeurs que peut prendre le taux de passage par zéro dans deux cas distincts.

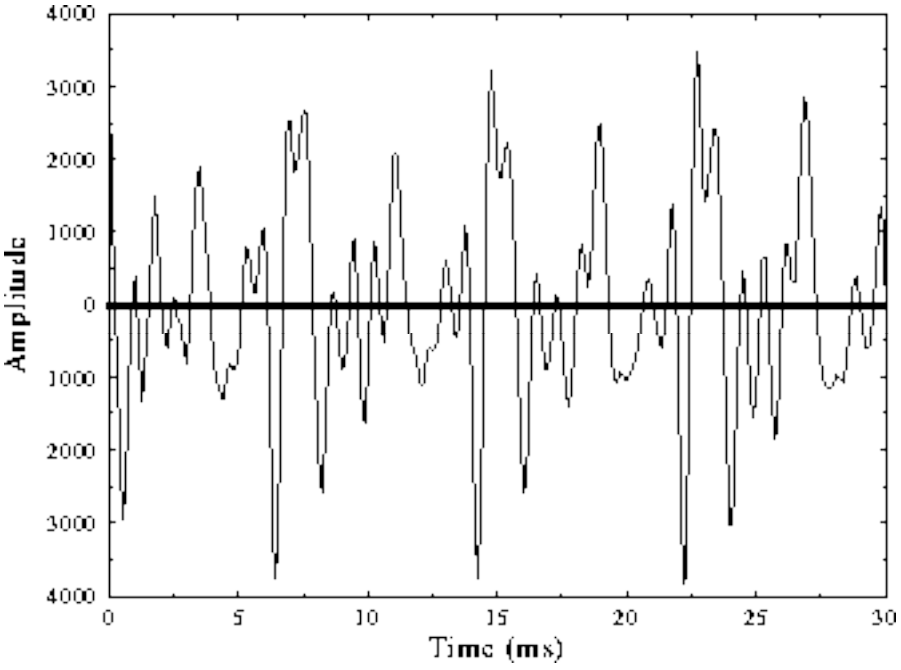


FIG. 8 – Exemple du taux de passage par zéro (faible) pour une zone voisée

3.1.4 Notions d'acoustique musicale

Dans cette partie, nous allons introduire les différentes caractéristiques propres à un son. Ces caractéristiques sont utilisées et exploitées tant par les scientifiques que par les artistes. Il est nécessaire de comprendre ces attributs afin de pouvoir effectuer correctement certains traitements numériques sonores.

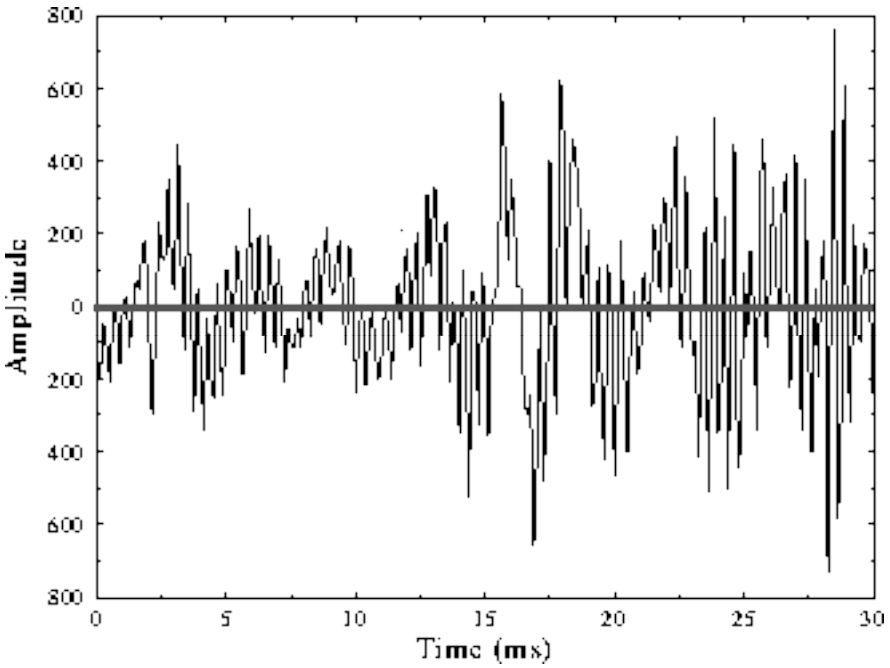


Fig. 9 – Exemple du taux de passage par zéro (fort) pour une zone non voisée

Amplitude

L'amplitude A du signal s est liée à son volume. En effet dans la section 3.1.5, nous présentons le principe de la perception de l'intensité sonore par le système auditif humain. L'amplitude d'un signal, s, s'exprime en dB et est définie par :

$$A_{dB}(s) = 20log\left(\frac{A(s)}{A_0}\right) \tag{61}$$

avec A_0 l'amplitude de référence (0dB) et $A(s)$ l'amplitude du signal s.

La dynamique d'un échantillon sonore est déterminée par les variantes d'amplitude des sources qui composent cet échantillon. Il existe plusieurs types de variations. Elles peuvent être linéaires dans le cas d'un fondu ou bien périodiques dans le cas de trémolo.

Timbre

Le timbre d'un son est en quelque sorte la couleur propre de ce son. Il varie en fonction de la source sonore. Du point de vue acoustique, le timbre est une notion très complexe qui dépend de la corrélation entre la fréquence fondamentale, et les harmoniques (ou partiels suivant leurs rapports avec la fréquence fondamentale). L'intensité respective de chaque harmonique est déterminante dans la caractérisation du timbre. Plus les fréquences de ces harmoniques sont proches des multiples entiers de la fréquence fondamentale, plus le son est pur ou harmonique (clavecin). Plus elles s'éloignent des multiples entiers, plus le son est in-harmonique (piano, cloche).

D'un point de vue musical, il conviendrait d'ajouter l'attaque du son, qui est d'une grande importance dans le message musical, en particulier en ce qui concerne l'articulation. L'étude de

l'attaque passe par celle des transitoires d'attaque, qui la caractérisent. Le timbre et l'attaque des sons nous permettent par exemple d'identifier sans le voir un instrument de musique quelconque, ou encore, de reconnaître au téléphone la voix d'une personne familière avant que celle-ci ne se soit présentée. L'expérience de l'audition en laboratoire d'un son dont l'attaque est supprimée montre que l'on devient totalement incapable de reconnaître la plupart des instruments de musique.

Partiels

Un son harmonique est très généralement composé d'une fréquence fondamentale, F_0, de plusieurs harmoniques, F_k, qui sont des fréquences multiples de F_0 : $\forall k,\ F_k = kF_0$. La famille des sons harmoniques se compose, notamment, des instruments de musique à vents, à corde, de la voix chantée ou parlée.

Dans certains cas, le principe d'harmonicité n'est pas exactement respecté ce qui signifie que les fréquences des harmoniques ne sont pas exactement multiples de la fondamentale. Dans la mesure où les écarts restent faibles il est tout de même envisageable de considérer le son comme harmonique, nous qualifions ces sons de *quasi-harmoniques*. Les sons quasi-harmoniques peuvent être obtenus avec certains instruments de musique percussifs ou à corde.

La définition du partiel est fortement liée à l'adverbe *lentement*. Il est possible de préciser la vitesse des variations. En effet, un son x composé d'un partiel dont l'amplitude est modulée par une sinusoïde, est équivalent à deux sinusoïdes :

$$\begin{aligned} x(t) &= \sin(2\pi ft)\sin(2\pi Ft+\theta) \\ &= \frac{1}{2}\cos(2\pi(F+f)t+\theta) - \frac{1}{2}\cos(2\pi(F-f)t+\theta) \\ &= \frac{1}{2}\sin\left(2\pi(F+f)t+\theta+\frac{\pi}{2}\right) + \frac{1}{2}\sin\left(2\pi(F-f)t+\theta-\frac{\pi}{2}\right) \end{aligned} \tag{62}$$

Ainsi, un partiel est défini comme une sinusoïde dont la fréquence et l'amplitude varient lentement au cours du temps. Dans le cas des sons harmoniques et quasi-harmoniques les partiels représentent la fondamentale et les harmoniques associées. De manière générale, les partiels sont les éléments du spectre d'un son obtenu par la transformée de Fourier.

Transitoires

Les transitoires sont assimilées à tout type de variations brusques temporelles d'amplitude dans un signal. Elles se rapportent notamment aux attaques qui désignent la très courte durée durant laquelle le son monte en amplitude avant d'atteindre sa valeur maximale. Cette période correspond à une hausse rapide de l'énergie dans tout le spectre. Il est possible de décomposer une transitoire en quatre grandes étapes : l'attaque, le déclin, le soutien et le relâchement. La figure 10 illustre cette première phase d'une transitoire. L'attaque peut parfois être plus douce, comme le montre la figure 11. Le déclin correspond à la période comprise entre le point fort de l'attaque et le moment où le son commence à conserver une amplitude constante. Le soutien est la phase durant laquelle le son est conservé à un niveau constant jusqu'à son relâchement. Enfin, le relâchement correspond à la phase de baisse d'intensité allant jusqu'à l'extinction total du volume (silence).

Le premier exemple, donné par la figure 10, représente une attaque très brusque suivi d'une extinction rapide comme le pincement d'une corde d'un clavecin.

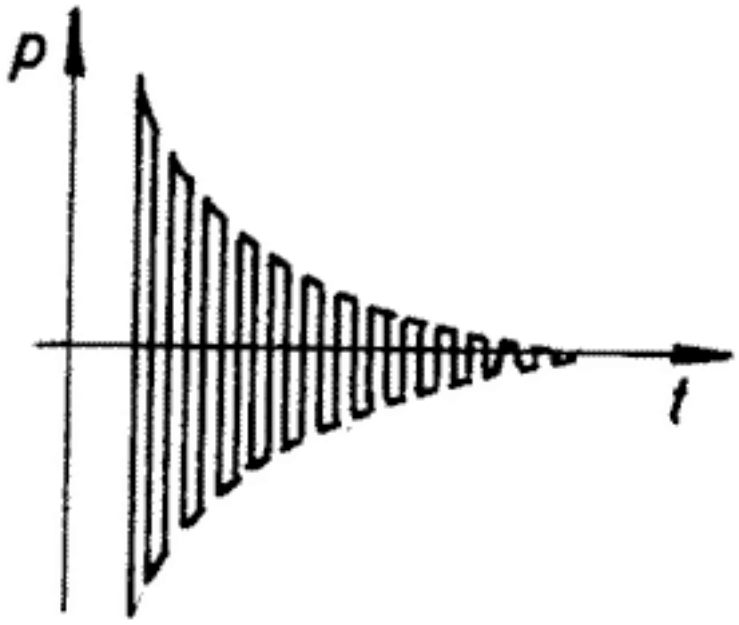


Fig. 10 – Représentation de la puissance du signal en fonction du temps d'un exemple d'attaque brusque : note de clavecin

Le deuxième exemple, donné par la figure 11, représente l'établissement progressif du son, note tenue puis extinction brusque comme un orgue à vent.

Les transitoires ont une importance pratiquement aussi grande que les harmoniques pour caractériser une source sonore. Il est donc important de savoir de quelle manière ces transitoires vont traverser la chaîne acoustique. Un fabricant d'appareils peut donner la rapidité avec laquelle un signal passe d'une tension à une autre, en indiquant la pente, S, exprimée généralement en microsecondes par volt.

D'autres appareils électroniques comme les samplings ou certains programmes informatiques permettent une programmation des transitoires, en manipulant des microsecondes.

3.1.5 Psychoacoustique

Dans cette section, nous définissons les principaux éléments et résultats dans le domaine de la perception. Les études menées en psychoacoustique permettent de définir des modèles simplifiés de perception suffisamment cohérents avec le mode de fonctionnement très complexe de l'oreille humaine. Certains travaux [Har97, ZF99] proposent des études de cas en psychoacoustique permettant de définir d'autres modèles plus complexes.

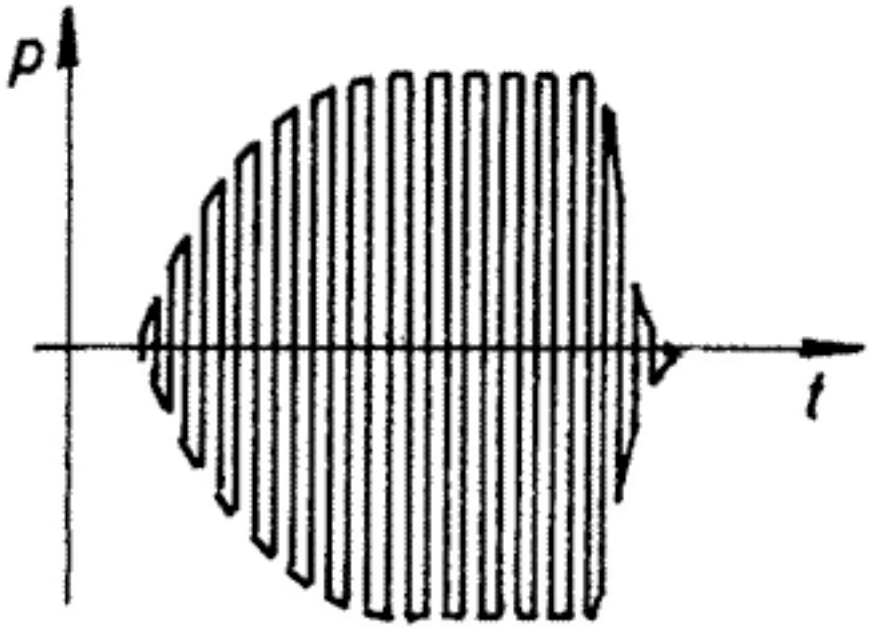


Fig. 11 – Représentation de la puissance du signal en fonction du temps d'un exemple d'établissement progressif : note d'orgue à vent

Perception de l'intensité

L'intensité d'un son, qu'il est aussi possible de nommer *force* ou encore *dynamique*, est la caractéristique nous permettant de distinguer un son fort d'un son faible (les musiciens parlent de nuances). Il s'agit, en termes scientifiques, de l'amplitude de la vibration, qui se mesure en décibels (dB). Dans cette dimension, comme dans celle de la hauteur, un son doit avoir une durée suffisamment longue pour que son intensité soit correctement appréciée. En effet, la sensation ne s'établit que progressivement et un son très bref, même s'il est intense, est perçu comme faible. Ce n'est que si sa durée se prolonge que la *sonie* croît. La sonie est une mesure de la perception acoustique de l'intensité. Elle permet de mesurer l'intensité des sons telle qu'elle est perçue chez l'homme. Elle est exprimée en phones ou en sones. Le seuil de perception pour que la sonie corresponde à l'intensité réelle, dans ce cas de figure, est mesuré au 1/10ème de seconde. Ce seuil explique certains phénomènes appelés effets de masquage, puisqu'un son intense, perçu rapidement, peut en masquer un autre qui l'a précédé, ou tout au moins en diminuer le niveau apparent.

La différence entre l'intensité réelle d'un son et l'intensité perçue par l'oreille humaine est expliquée par le fonctionnement du système auditif qui tient compte de nombreux facteurs temporels et fréquentiels. En effet, l'oreille n'est pas équitablement sensible aux fréquences, la figure 12 montre les courbes de même sonie en fonction de la fréquence (courbes de Fletcher-Munson). De plus, des études psychoacoustiques ont démontré que la perception de l'intensité par l'oreille n'est pas linéaire mais plutôt logarithmique. Pour cela, l'échelle décibel (dB) a été définie par :

$$L(x) = 20log\left(\frac{P(x)}{P_0}\right) \tag{63}$$

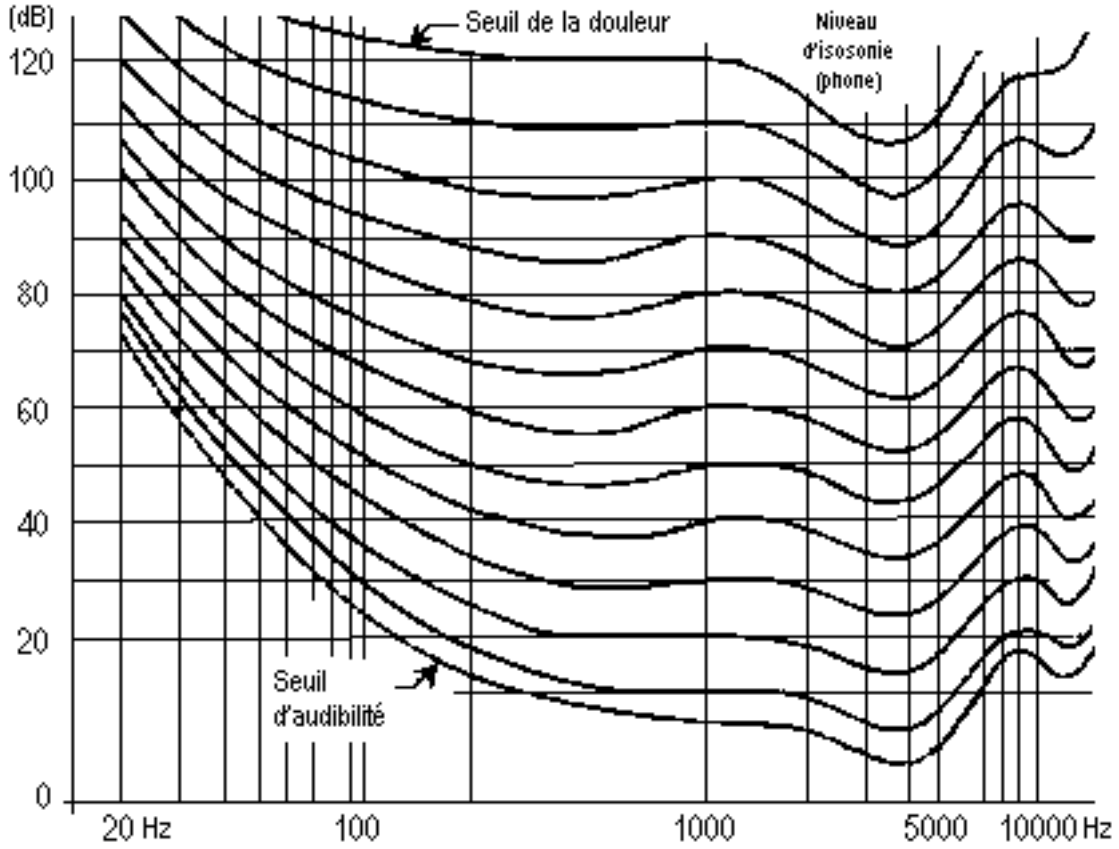


Fig. 12 – Courbes d'intensité perçue en fonction de la fréquence (Courbes de Fletcher-Munson 1993)

avec $L(x)$ le niveau sonore du signal x en dB, $P(x)$ la pression du signal x qui est égale au carré de l'intensité et P_0 la pression acoustique de référence (10^{-5} Pa) correspondant au seuil de l'audition à 1000Hz. L'équation précédente (63) peut-être ré-écrite en fonction de l'amplitude, $A(x)$, du signal puisque celle-ci est proportionnelle à la pression acoustique. D'où :

$$L(x) = 20log\left(\frac{A(x)}{A_0}\right) \tag{64}$$

Enfin, la notion de volume d'un son est intimement liée au niveau sonore L (voir équations 63 et 64) à tel point qu'il est possible d'assimiler ces deux grandeurs égales.

L'énergie

L'énergie d'un signal, exprimée comme RMS (Root Mean Square), est une mesure de l'amplitude de ce signal. Dans le cas discret, la RMS est définie par :

$$RMS(x) = \sqrt{\frac{1}{W_s}\sum_{n=1}^{W_s} x^2[n]} \tag{65}$$

avec W_s le nombre total d'échantillons du signal x.

Perception de la hauteur

La hauteur est un paramètre perceptif et musical relatif au paramètre physique qu'est la fréquence. L'étude de la perception de la hauteur par le système auditif humain est très complexe car beaucoup de paramètres entrent en jeu.

Dans le cas d'un son harmonique ou quasi-harmonique, le système auditif humain ne perçoit qu'une seule hauteur qui correspond généralement à la fréquence de la fondamentale. Cependant, il peut arriver que la hauteur perçue ne corresponde à aucune fréquence présente. Dans ce cas-là, il s'agit du phénomène de la fondamentale manquante.

Enfin, bien que le système auditif soit sensible à une large bande de fréquences variant entre 20Hz et 20000Hz, il en de même vis-à-vis des hauteurs [Roa98].

Perception de la fréquence

Comme brièvement dit dans la section précédente, le système auditif humain est sensible aux fréquences variant de 20Hz à 20000Hz. Toutefois, l'étude de la perception de ces fréquences n'est pas chose facile. En effet, des expérimentations ont démontré que lorsque deux fréquences sont trop éloignées, le système auditif arrive à les discerner. En revanche, si elles sont trop proches le cerveau a du mal à les discerner. Nous sommes alors en présence du phénomène de *fusion* ce qui donne une impression de battements [Pre00]. En d'autres termes une seule fréquence qui varie est alors perçue. La différence de valeur des fréquences permettant de pouvoir les discerner n'est pas linéaire et dépend aussi des valeurs de ces fréquences.

Bandes de Barks

Les bandes de Barks sont une échelle de mesure utilisée dans le domaine fréquentiel [ZF99]. Il s'agit d'une échelle non-linéaire. En effet, en dessous de 500 Hz, la largeur des bandes critiques est constante. Au dessus de 500Hz, cette largeur augmente avec la fréquence. Il existe des formules qui fournissent une bonne approximation pour la conversion des Hertz en Barks. Soit b la bande de Bark et f une fréquence en Hz alors :

$$b(f) = \begin{cases} \frac{f}{100} & \text{pour } f \leq 500 \\ 9 + 4log_2\left(\frac{f}{1000}\right) & \text{pour } f > 500 \end{cases} \quad (66)$$

$$f(b) = \begin{cases} 100b & \text{pour } b \leq 5 \\ 1000 \times 2^{\frac{b-9}{4}} & \text{pour } b > 5 \end{cases} \quad (67)$$

Il est aussi possible d'obtenir la largeur d'une bande critique en fonction de la fréquence centrale :

$$\Delta f = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69} \quad (68)$$

avec Δf la largeur de la bande critique en Hz et f la fréquence centrale en Hz également.

Échelle Mel

Utilisée par certains auteurs [PSAO02a] dans la littérature scientifique, *l'échelle Mel* tente de mettre en relation la tonie et les hauteurs perçues. Elle correspond à des emplacements de hauteurs sur le continu qui va de grave à aiguë. Elle n'établit pas une correspondance entre les fréquences et les mels mais tente d'établir un lien entre les sensations de hauteurs perçues. Elle met principalement une chose en valeur, c'est que sur le plan de la perception, une octave entre deux sons graves, 250 et 500 Hz par exemple (250 mels), paraît plus petite qu'une octave entre deux sons aigus, 1000 et 2000 Hz (1800 mels).

3.1.6 Analyse dans le domaine de la parole et de la musique

De nombreux travaux ont été consacrés à l'analyse de la parole et de la musique. On doit à Rabiner [RJ93] un ensemble de travaux novateurs et originaux sur l'étude de la parole. Dans [TC99], l'auteur propose un ensemble de descripteurs audio bas niveau pour la navigation rapide dans les grandes bases de données audio. Pour cela, l'auteur utilise les descripteurs suivants : le centroïde spectral, le roll-off spectral, le taux de passage par zéro, l'énergie, la moyenne et la variance associées à chacun des descripteurs précédemment énoncés. L'approche présentée dans [AO98] consiste à segmenter les bandes sonores contenant de la parole à l'aide d'une méthode statistique.

Toutefois, dans [MB03], les auteurs proposent un ensemble de descripteurs pour la classification audio mais aussi de la musique. Ils proposent de caractériser différents styles musicaux : Musique Classique et Musique Populaire. Pour cela les auteurs font appel à un descripteur très utilisé dans le domaine de la parole mais qui peut être adapté au domaine musical : les MFCC (Mel-Frequency Cepstral Coefficients). Ils ne considèrent que les treize premiers coefficients. Dans un autre registre, les auteurs de [GPD00] proposent une méthode de classification des sons percussifs à l'aide de descripteurs comme notamment le taux de passage par zéro. L'originalité des travaux présentés dans [ZP04] réside dans la méthode d'extraction des descripteurs. En effet, il est proposé une méthode automatique d'extraction des descripteurs musicaux en fonction du problème posé. Les travaux présentés dans [LPR03] ont pour but de proposer une classification des instruments présents dans un morceau de musique. Concernant le choix des descripteurs, les auteurs ont utilisé ceux présentés dans [GR02]. Ils sont au nombre de 162, dont notamment les MFCC et le taux de passage par zéro. Enfin, dans [TC02], l'auteur propose une approche pour la classification des genres de musique. L'auteur utilise des descripteurs tels que les MFCC, le spectre du signal obtenu par transformation de Fourier, le roll-off spectral, le taux de passage par zéro.

Les travaux [Foo97, DW99, DW00] utilisent les coefficients MFC pour leurs travaux d'indexation de la bande sonore (voir section 3.2). De même, dans les travaux de [PSAO02a, PRMAO03] des sous-systèmes de classifications sont proposés : Parole/NonParole et Musique/NonMusique basés sur les coefficients cepstraux. D'autres travaux ont été menés dans l'indexation de la parole et de la musique. De plus amples détails concernant ces travaux sont apportés dans la section suivante. Dans [PT04], les auteurs proposent un système d'indexation de la bande sonore en Parole/Musique uniquement basé sur l'énergie (RMS) et sur le taux de passage par zéro. Dans le même registre, [EMKPK00] propose une méthode d'indexation de la bande sonore en Parole/Musique basée aussi sur le taux de passage par zéro. Enfin, les travaux de [PSAO02b] sont basés sur l'utilisation du centroïde spectral et du taux de passage

pas zéro pour l'indexation en Parole/Musique. Les descripteurs utilisés dans ces travaux sont injectés dans un système pour caractériser les segments de parole et de musique du flux audio.

3.1.7 Analyse dans le domaine du silence

Comme dit précédemment, les domaines audio les plus explorés sont la musique et la parole (voir section précédente). Toutefois, quelques travaux ont été proposés pour la détection des silences et l'étude des bruits (voir section suivante). Cependant, la détection des silences n'est pas une fin en soi car elle est souvent utilisée dans les études liées à la parole [AR76, LRRW81, Ney81, RS77].

Un des travaux dans le domaine est présenté dans [Sou83]. Bien que relativement anciens, les travaux proposés donnent de très bons résultats qui rivalisent parfaitement avec les travaux actuels [JEA99, WR00]. L'auteur propose une méthode de détection statistique à partir de descripteurs tels que ceux qui sont présentés dans [RS75] dont notamment l'énergie, les MFCC, le taux de passage par zéro et le coefficient d'autocorrélation. Le coefficient d'autocorrélation n'est autre que le rapport entre les valeurs des deux premiers maxima de la fonction d'autocorrélation.

Des travaux plus récents sur la détection des silences ont vu le jour notamment ceux de [BCC99]. Ces travaux s'inscrivent dans un système de segmentation Parole/Silence de la bande audio. Parmi les descripteurs utilisés il y a le taux de passage par zéro très utilisé dans le domaine de la parole. D'autres travaux de la même génération [JEA99, WR00], proposent un détecteur de silence dans le domaine du multimédia. Dans [JEA99], la méthode utilisée est uniquement basée sur l'énergie du signal calculée dans le domaine logarithmique. Enfin dans [WR00], la méthode proposée recherche les consonnes et les voyelles dans les segments de parole des bandes audio. Les silences sont détectés comme étant des zones qui ne contiennent pas de parole, sachant que les auteurs ne traitent que des extraits de parole. Les silences détectés sont donc les silences entre les mots et les locuteurs.

3.1.8 Analyse dans le domaine du bruit

La classe des bruits a suscité peu d'intérêt dans la domaine de la recherche. Toutefois, les travaux sur les bruits (comme pour les silences) sont souvent intégrés dans un classifieur Parole/Musique/Bruit/Silence mais ne représentent pas une fin en soi.

Dans les travaux présentés dans [DBAF99, ZK99, LJZ01], les différents auteurs proposent des méthodes d'indexation de la bande sonore qui prennent en compte la classe des bruits. Ces méthodes d'indexation seront développées en détail dans la section 3.2. En ce qui concerne les descripteurs utilisés dans ces approches ils se rapportent à ceux couramment considérés pour l'étude de la musique, de la parole et des silences. Il s'agit donc de descripteurs tels que le taux de passage par zéro, les coefficients MFC, le centroïde spectral, le roll-off spectral, etc... En effet, à la suite de multiples travaux, les descripteurs énoncés ci-dessus se sont révélés être de bons candidats pour caractériser la parole, la musique et les silences. De plus, ces descripteurs, de part leurs propriétés, permettent de différencier les différentes classes audio entre elles.

Cependant dans [DBAP00], les auteurs proposent une méthode pour reconnaître un ensemble de bruits qu'ils ont préalablement défini. Parmi ces bruits, il y a des claquements de

porte, des bris de glace, des cris humains, des explosions, des coups de pistolet, etc...Les auteurs ont fait appel à un descripteur qui se compose de plusieurs composantes (N), il s'agit de la dérivée de l'énergie des N bandes spectrales du signal. Pour obtenir ce descripteur, il faut tout d'abord récupérer le spectre du signal, obtenu par transformée de Fourier, puis découper le spectre en N bandes de fréquences de tailles fixes et enfin calculer l'énergie du spectre pour chacune de ces bandes.

Enfin, un des challenges dans le domaine de l'analyse et de l'indexation consiste à proposer des descripteurs audio les mieux adaptés. Dans [BPJ02] les auteurs proposent une méthode pour extraire des descripteurs pertinents pour l'étude des bruits. Cette méthode s'appelle *Distortion Discriminant Analysis* (DDA), elle s'apparente un peu à la méthode présentée dans [ZP04] concernant l'extraction automatique de descripteurs dans le domaine de la musique. La méthode DDA utilise le principe de l'Analyse par Composante Principale (PCA). Les résultats obtenus après différentes expérimentations montrent que la DDA permet de gérer des descripteurs pertinents pour l'étude des bruits.

3.2 Indexation audio

Dans cette partie, nous présentons une vue d'ensemble des différentes applications de l'analyse audio, présentée dans la section précédente. Les principaux travaux en indexation audio consistent à caractériser les quatre grandes classes audio dans les bandes son de tout type de contenu, comme par exemple dans [ZK01, LJZ01]. Toutefois, certains travaux comme [TC04] s'intéressent à des types particuliers de contenus audio. Il est important de préciser que dans les parties qui suivent seuls certains travaux du domaine sont présentés. En effet, le nombre important de travaux associés à ce domaine publiés dans la littérature suggèrent une sélection arbitraire car il semble très difficile d'en faire une liste complètement exhaustive.

3.2.1 Classification audio de contenus génériques

Dans cette partie, nous présentons différents travaux de recherche dans le domaine de la classification audio. Ces dernières années, le monde de la recherche a vu émerger ce domaine de recherche qui constitue une part importante des applications directement liées aux études menées dans le domaine de l'analyse audio. Ainsi, les sous-parties qui suivent sont consacrées à chacun des travaux que nous avons sélectionnés parmi l'ensemble des travaux présents dans la littérature.

Zhang et Kuo [ZK01]

Dans ces travaux [ZK01], les auteurs proposent une méthode de segmentation temps réel de la bande sonore des programmes télévisuels. Cette segmentation consiste à caractériser différentes classes audio telles que : parole, musique, chanson, son d'ambiance, parole avec musique de fond, son d'ambiance avec musique de fond, silence, etc ...

Pour satisfaire la contrainte du traitement temps réel, les auteurs proposent d'utiliser des descripteurs simples tels que l'énergie, le taux de passage par zéro, l'extraction de la fondamentale. Leur travail se situe dans la lignée des travaux précédents de [Sau96, KW96].

Dans [KW96], la méthode proposée consistait en la segmentation selon parole, silence, rires et non parole. Les descripteurs utilisés sont les coefficients MFC et les modèles de Markov cachés pour la règle de décision. Enfin, concernant l'approche proposée par [PFE96] dont le but est de caractériser les segments qui indiquent de la violence au niveau du flux audio : cri, tir de pistolet, explosion.

Pour ce qui est de l'analyse audio, les auteurs font appel à l'énergie à court terme définie par :

$$E_n = \frac{1}{N}\sum_m (x[m]w[n-m])^2 \tag{69}$$

avec $x[m]$ le signal audio sous sa forme discrète, n l'indice temporel de l'énergie à court-terme et $w[m]$ une fenêtre rectangulaire de longueur N.
Dans ces travaux, ce descripteur a été utilisé pour séparer les parties voisées et non voisées de la parole, pour détecter les zones de silence et enfin pour caractériser les changements de style de musique.
Parmi les autres descripteurs utilisés les auteurs utilisent le taux de passage par zéro à court terme défini par :

$$Z_n = \frac{1}{2}\sum_m |sgn[x[m]] - sgn[x[m-1]]w[n-m] \tag{70}$$

avec $w[n]$ une fenêtre rectangulaire et

$$sgn[x[n]] = \begin{cases} 1, & x[n] \geq 0, \\ -1, & x[n] < 0 \end{cases}$$

Ce descripteur est un bon candidat pour la caractérisation des parties voisées et non voisées de la parole ainsi que pour la classification des sons d'ambiance car la forme de la courbe des valeurs du taux de passage par zéro est différente en terme de régularité, périodicité, ... suivant le type de son d'ambiance rencontré.
Enfin le dernier descripteur est la détection de la fondamentale utilisé bien évidemment dans le domaine musical. Ce descripteur est extrait à partir du spectre du signal. La méthode utilisée pour la détection de la fondamentale consiste, dans un premier temps, à localiser les maxima du spectre. Puis, dans un second temps, l'algorithme tente de mettre en avant un diviseur commun pour l'ensemble des fréquences correspondantes aux maxima précédemment détectés dans le spectre. Si un diviseur commun n'est pas trouvé alors la fréquence de la fondamentale est fixée à zéro.

L'indexation audio se décompose en deux grandes parties :

- La détection des frontières des différents segments audio et
- La classification de chacun des segments précédemment détectés.

Un des points forts de ces travaux est la contrainte temps réel fixée par les auteurs. Les auteurs extraient et analysent les descripteurs à la volée sur les données audio entrantes. Chaque changement abrupt des valeurs de n'importe quel descripteur est considéré comme une frontière de segment audio. De plus, la figure 13 ci-dessous illustre le principe des fenêtres glissantes utilisé pour étudier les variations de la forme des courbes des descripteurs. Dans chacune des deux fenêtres, $win1$ et $win2$, les moyennes des valeurs sont calculées et sont notées $Ave(w1)$ et $Ave(w2)$ pour les fenêtres $win1$ et $win2$ respectivement. Ainsi, pour chaque nouvelle valeur de descripteur calculée, l'ensemble est mis à jour en temps réel. Les valeurs

$Ave(w1)$ et $Ave(w2)$ sont comparées. Si une trop grande différence entre elles est observée alors le point correspondant au coté commun aux deux fenêtres, noté E sur la figure 13, est considéré comme une frontière de segment audio également. Dans leurs expérimentations les auteurs ont utilisé une taille de fenêtre équivalente à 100 valeurs de descripteurs. Après avoir

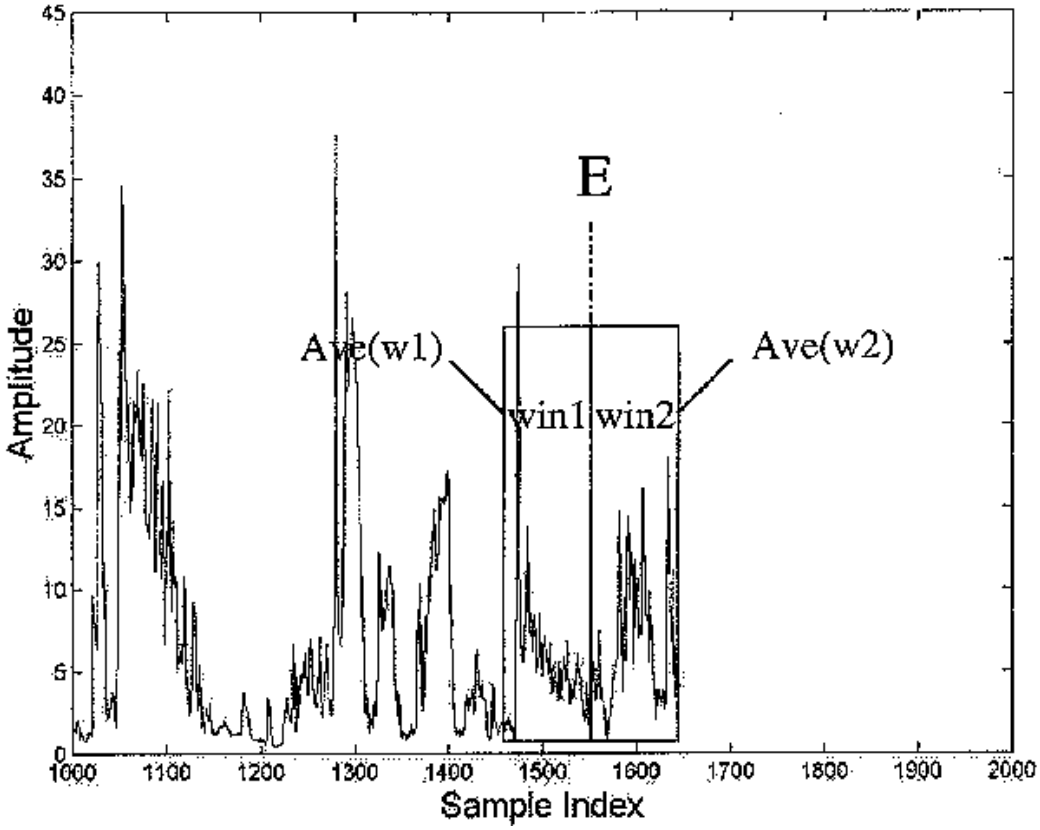


Fig. 13 – Principe de la fenêtre glissante pour l'indexation audio [ZK01]

détecté les frontières de classes audio, l'étape suivante consiste à étiqueter ces segments audio suivant les classes prédéfinies au départ.
La première étape consiste à détecter les silences. Un silence est défini comme étant une partie comprenant un signal audio non perceptible par l'oreille humaine. Les auteurs utilisent la combinaison de l'énergie et du taux de passage par zéro. En effet, seule l'énergie n'est pas suffisamment discriminante dans le cas de segments musicaux ou de bruit de faible intensité. Si les valeurs des courbes de l'énergie et du taux de passage par zéro sont inférieures à un seuil alors le segment audio est étiqueté comme un silence.
La seconde étape a pour but de caractériser les segments audio contenant une composante musicale. Seuls les segments qui n'ont pas été étiquetés comme silence lors de l'étape précédente sont considérés. La méthode utilisée ici consiste à rechercher une série de pics de fréquence stable et continue dans le spectre du signal. Afin de s'affranchir des problèmes de sur-détection liés à la présence d'harmoniques dans la parole et dans les basses fréquences des bruits, seules les fréquences au-delà de 500Hz sont considérées. De même les signaux dont l'énergie est trop faible sont ignorés. Une valeur booléenne (0 ou 1) est affectée pour les segments restants qui contiennent, ou non, une composante musicale. La proportion des valeurs 0 affectées est

ensuite calculée en fonction du nombre total de segments classifiés, les auteurs appellent cette proportion le *zéro ratio*. Plus cette proportion est élevée, moins le signal considéré contient une composante musicale. De surcroît, l'analyse de cette proportion permet aussi de dégager les propriétés suivantes :

- *Parole* : Le zéro ratio est tout le temps supérieur à 0.95.
- *Sons d'ambiance* : Les sons d'environnement harmoniques sont considérés comme contenant une composante musicale, contrairement au cas des sons non harmoniques.
- *Musique pure* : Le zéro ratio dans le cas de la musique pure est en dessous de 0.3.
- *Chansons* : La plupart des segments contenant de la chanson ont un zéro ratio inférieur à 0.5.
- *Parole avec musique de fond* : Naturellement si l'énergie de la composante parole est beaucoup plus élevée que celle de la composante musique du signal alors la musique de fond est masquée par la parole et ne peut pas être détectée. Dans le cas contraire deux cas sont envisageables. L'énergie de la composante musique est supérieure à celle de la parole, dans ce cas la valeur du zéro ratio est inférieure à 0.6. Dans le cas où l'énergie de la composante parole est supérieure alors la valeur du zéro ratio est supérieure à 0.8.

Ainsi avec un seuil fixé à 0.7 et l'emploi de certaines règles, les auteurs affirment être en mesure de caractériser les segments contenant une composante musicale.
La troisième phase concerne la détection des sons d'ambiance harmoniques. Pour cela, les auteurs analysent le spectre du signal afin d'extraire et de caractériser une fréquence fondamentale ainsi que les harmoniques associées. Si c'est le cas alors le segment considéré est étiqueté comme tel.
La quatrième étape consiste à distinguer les segments contenant uniquement de la musique. La méthode proposée est basée sur l'analyse des valeurs du taux de passage par zéro et sur la détection de la fondamentale.
L'étape suivante consiste à distinguer les segments contenant uniquement de la parole. Les auteurs font appel au taux de passage par zéro, à la détection de la fondamentale ainsi qu'à l'énergie. Dans un premier temps, les auteurs s'intéressent à la relation qu'il existe entre l'énergie et le taux de passage par zéro dans le domaine de la parole. En effet, dans les parties non voisées la courbe des valeurs du taux de passage par zéro présente des pics alors que la courbe de l'énergie forme des creux. Au contraire, dans le cas des parties voisées c'est la courbe de l'énergie qui présente des pics et celle du taux de passage par zéro qui contient des creux. Dans un deuxième temps, les auteurs calculent la moyenne et la variance du taux de passage par zéro pour chaque période de temps donnée. Enfin les auteurs recherchent la présence d'harmonicité dans le spectre, car la parole est un mélange d'harmonicité et de non harmonicité, il y a donc un certain pourcentage d'harmonicité que tente de caractériser les auteurs. De la même manière que pour la détection de la musique pure, les auteurs comparent les différentes valeurs de leurs descripteurs avec des seuils et une valeur de décision comprise entre 0 et 1 fixée de la même manière que précédemment. Si cette valeur est supérieure à 0.5 alors le segment considéré est classifié comme étant composé purement de parole.
Enfin la dernière phase a pour but de proposer une classification des sons d'ambiance non harmoniques. Les segments considérés sont ceux n'ayant pas été classifiés précédemment. Les auteurs proposent quatre classes de sons d'ambiance définis par :

- Périodiques ou quasi-périodiques : Si la courbe des valeurs de l'énergie ou du taux de passage par zéro admet des pics à intervalles réguliers ou quasi-réguliers.

- Mélange harmoniques et non harmoniques : Si la proportion d'harmonicité est égale à un certain seuil (plus petit que la musique mais plus grand que pour sons non harmoniques).
- Non harmonique et stable : Si les valeurs du taux de passage par zéro sont comprises dans un petit intervalle par rapport au taux d'harmonicité.
- Non harmonique et irrégulier : Tout ce qui ne satisfait pas les trois précédents cas.

Les expérimentations ont été menées sur une base de données d'environ 1000 sons. Chaque extrait a une durée variant de quelques secondes à 1 minute. Les auteurs ont aussi composé leur base de test avec des extraits de bande sonore de programmes audiovisuels de tout type (film, séries, documentaires, etc...). Ainsi, la base finale est composée de 1200 sons au total. Les performances de la classification sur l'ensemble des sons de la base de test décrite ci-dessus sont données par le tableau de la figure 14. Les résultats démontrent que la méthode proposée

Classe Audio	Rappel	Précision
Silence	100%	100%
Avec une composante musicale	100%	95.3%
Sans une composante musicale	97.8%	100%
Pure Parole	91.9%	91%
Pure Musique	97.9%	94.5%
Sons d'ambiance non harmoniques	95.3%	98.5%

Fig. 14 – Résultats de la classification audio [ZK01]

est très robuste. Elle a été testée sur un grand ensemble de sons très variés et les valeurs des performances sont très fortes. De surcroît, cette méthode est temps réel ce qui implique qu'elle fait appel à des descripteurs dont le calcul est peu gourmand en terme de ressources. Ainsi, tout réside dans la méthode d'analyse des descripteurs et leur interprétation en fonction du contexte.

Lu, Jiang et Zhang [LJZ01]

Ce travail a comme objectif la classification et la segmentation des bandes sonores. Les classes définies par les auteurs sont parole, musique, bruit et silence. La méthode présentée se décompose en deux grandes étapes. La première d'entre elles consiste à séparer les segments parole des segments de non parole. La seconde étape consiste à sous-classifier les segments non parole selon musique, bruit et silence. Pour cela les auteurs s'inspirent de nombreux travaux existants comme [WBW96, Foo97].

Dans ces travaux de segmentation les auteurs utilisent le flux spectral et les variantes de descripteurs connus : le taux de passage par zéro et l'énergie. De plus, la périodicité de bande et la proportion de fenêtres bruitées du signal sont aussi utilisées.
La première de ces variantes est la proportion haute du taux de passage par zéro (HZCRR). Il s'agit de la proportion de fenêtres d'analyse dont la valeur du taux de passage par zéro est supérieure à 1.5 fois la valeur moyenne du taux de passage par zéro sur une durée d'une

seconde. Le HZCRR est défini formellement comme :

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(ZCR(n) - 1.5avZCR) + 1] \quad (71)$$

$$avZCR = \frac{1}{N}\sum_{n=0}^{N-1} ZCR(n) \quad (72)$$

avec n l'indice de la fenêtre d'analyse, N le nombre total de fenêtres d'analyse équivalent à 1 seconde, $sgn[.]$ la fonction qui renvoie 1 ou -1 suivant le signe et $ZCR(n)$ la valeur du taux de passage par zéro pour la fenêtre d'indice n.
Ce descripteur va être utilisé pour la distinction entre parole et musique, en effet les valeurs du HZCRR seront plus élevées pour la parole que pour la musique. La raison provient de la structure de la parole, composée de parties voisées et non voisées, que n'a pas la musique.
La deuxième variante est la proportion basse d'énergie à court-terme (LSTER). Il s'agit de la proportion de fenêtres d'analyse dont l'énergie est 0.5 fois plus faible que la moyenne de l'énergie sur une durée de une seconde. La LSTER est définie formellement comme :

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[sgn(0.5avSTE - STE(n)) + 1] \quad (73)$$

$$avSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n) \quad (74)$$

avec N le nombre total de fenêtres d'analyse équivalent à 1 seconde, $STE(n)$ l'énergie pour la fenêtre d'indice n et $avSTE$ la valeur moyenne de la STE sur une durée d'une seconde.
La LSTER est utilisée en complément du HZCRR pour la distinction de la parole et de la musique mais aussi pour la détection des silences.
Les auteurs utilisent aussi le flux spectral déjà défini dans la section précédente.
La périodicité de bande est définie comme la périodicité de chaque sous-bande du spectre. Les sous-bandes définies sont : 500-1000Hz, 1000-2000Hz, 2000-3000Hz et 3000-4000Hz. La périodicité de chaque sous-bande peut-être définie par l'équation suivante :

$$bp_i = \frac{1}{N}\sum_{j=0}^{N-1} max(r_{i,j}(k_p), r_{i,j}(k_p) - c), \forall i = 1, \ldots, 4 \quad (75)$$

avec bp_i la valeur de la périodicité de la sous-bande i, N le nombre total de fenêtres d'analyse dans un clip audio, c un paramètre utilisé pour éliminer l'effet de lissage dû au filtrage passe-bas [MB95], $r_{i,j}(k_p)$ le maximum local de la fonction d'autocorrélation normalisée pour la sous-bande d'indice i de la fenêtre d'analyse d'indice j et k_p l'index du maximum local.
À titre d'exemple, la périodicité de bande vaut 1 pour une sinusoïde et 0 dans le cas d'un bruit blanc. Ce descripteur est un bon candidat pour la discrimination de la musique et du bruit. En effet, les valeurs de ce descripteur seront élevées dans le cas de la présence de musique et assez faibles dans le cas de bruits.
Le dernier descripteur utilisé dans ces travaux est la proportion de fenêtres bruitées. Une fenêtre d'analyse est considérée comme bruitée si le maximum local de sa fonction d'autocorrélation normalisée est inférieur à un certain seuil. Naturellement, les valeurs de ce descripteur

sont plus élevées en présence de bruit que de musique. Il sera donc utile pour la discrimination entre le bruit et la musique.

Dans leur démarche de classification, les auteurs séparent d'abord la parole du reste, puis ils classifient plus finement, en Musique, Bruit et Silence, les plages de non-parole. Le schéma de cette classification est présenté sur la figure 15. La première phase de l'indexation consiste

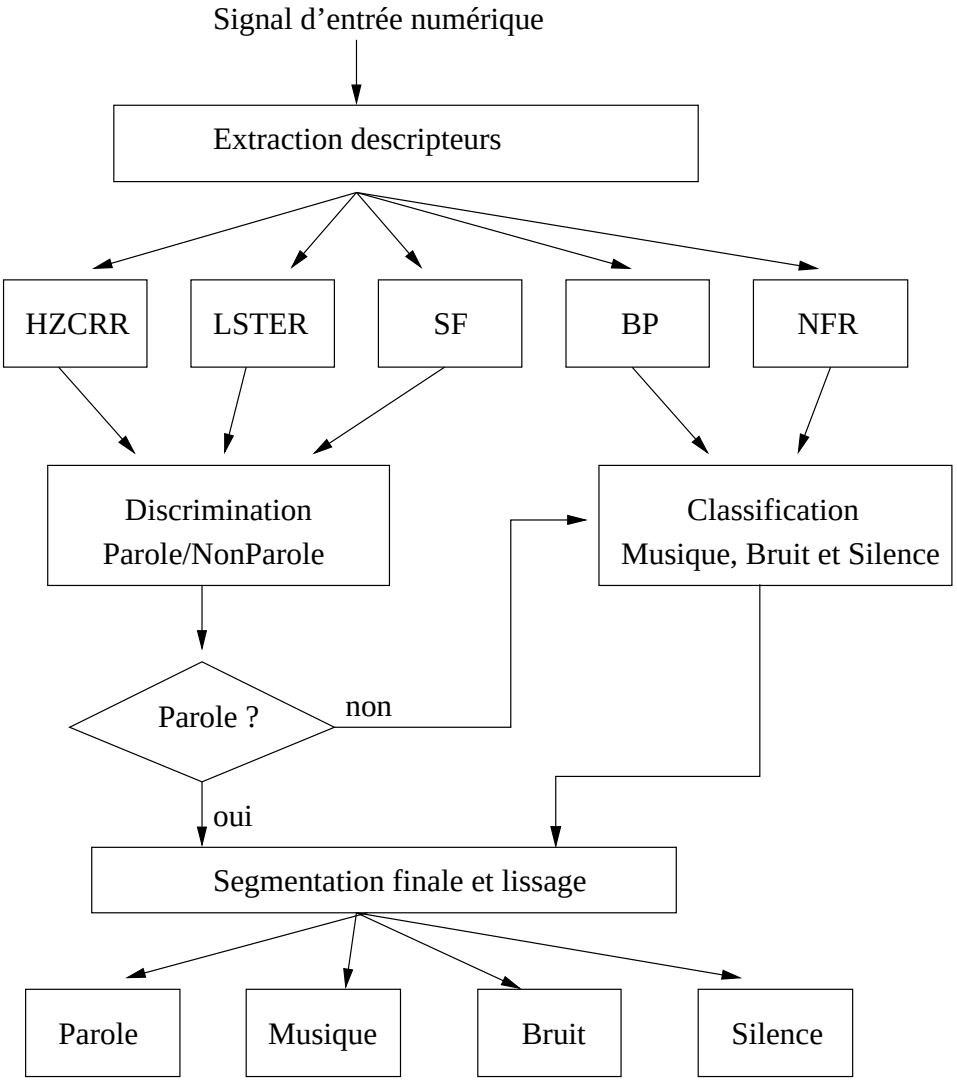


Fig. 15 – Schéma de fonctionnement de la méthode de classification et de segmentation audio [LJZ01]

à caractériser les segments de parole et de non parole. Les descripteurs utilisés sont donc le HZCRR, la LSTER et le SF (voir figure 15). Ces descripteurs sont utilisés car ils ont un fort pouvoir de discrimination et ne nécessitent pas de fortes ressources pour être extraits. Les auteurs utilisent une méthode de classification basée sur le principe du K-Nearest-Neighbor (KNN) [FH52] avec le paramètre $k = 2$ et ils supposent que les valeurs des descripteurs suivent le modèle gaussien. Le système de classification proposé est temps réel.
Dans la classification en Musique, Bruit et Silence (MBS), les auteurs utilisent de simples

règles de seuillage avec les seuils sur les valeurs des descripteurs appris expérimentalement.
La dernière phase du processus de segmentation, comme le montre la figure 15, consiste à affiner le résultat de la segmentation obtenu précédemment. Les auteurs supposent que le signal audio est continu et que par conséquent il est très peu probable que des changements brusques et fréquents puissent apparaître. Ainsi, les auteurs ont défini un certain nombre de règles qu'ils appliquent sur le flux audio de telle sorte que si un motif du type *parole-musique-parole* est détecté de manière récurrente sur plusieurs fenêtres consécutives alors ces fenêtres sont considérées comme étant uniquement de type *parole*.

Les expérimentations ont été menées sur une base de données comprenant des échantillons de parole dans différentes conditions, des échantillons de musique de différents genres ainsi que différents types de bruit.
La durée totale est de six heures, deux heures ont été allouées à l'apprentissage du système et quatre heures pour les tests. Les performances globales sont résumées par le tableau de la figure 16. Tout d'abord, il est dommage que les auteurs ne renseignent pas sur les

Classe Audio	Résultats de discrimination		
	Parole	Musique	Bruit
Parole	97.45%	1.55%	1.00%
Musique	3.16%	93.04%	3.80%
Bruit	10.49%	5.08%	84.43%

Fig. 16 – Résultats de la classification audio [LJZ01]

performances de la détection des silences. Concernant les résultats affichés, ils s'avèrent être très bons relativement à la richesse et à la taille des éléments du corpus de test. La classification des bruits reste en retrait par rapport à la musique qui est moins performante que la parole.

Autres méthodes

Bien d'autres méthodes existent dans la littérature dans ce domaine. Nous pouvons citer entre autres [ABL01, HCA01, LZJ02]. Ces travaux concernent la segmentation et l'indexation des flux audio à partir de descripteurs classiques tels que ceux que nous avons étudiés dans la section précédente 3.1. Quant aux méthodes de segmentation utilisées, elles sont identiques à celle présentées précédemment. Les différences résident dans l'utilisation de certains descripteurs inhabituels au domaine mais dont l'analyse apporte des informations pertinentes.

3.2.2 Caractérisation de scènes audio et application au traitement de la vidéo

Dans cette section nous présentons une autre application de l'indexation du flux audio. Il s'agit de modéliser des scènes audio et d'appliquer ce modèle au domaine de la vidéo. Pour cela, nous présentons, entre autres, les travaux de [VRSB99].

Venugopal, Ramakrishnan, Srinivas et Balakrishnan [VRSB99]

Dans ces travaux les auteurs proposent un modèle de scène audio comme étant une séquence de plans de montage vidéo durant lesquels les caractéristiques audio ne changent pas. Par ce modèle, les auteurs entendent vouloir donner un sens sémantique aux différentes scènes ainsi localisées.

Les auteurs proposent une méthode de segmentation audio en parole, musique et silence basée sur les travaux de [HW94, Sau96, SS97]. Notamment concernant les descripteurs utilisés :

- Tonalité : Caractérisation de l'harmonicité du spectre.
- Bande passante : La parole occupe une bande de fréquence réduite dans le spectre, jusqu'à environ 8kHz. Contrairement à la musique qui occupe tout le spectre jusqu'à 20kHz.
- Hauteur : La hauteur pour la parole peut être perçue sur seulement trois octaves, alors que dans le cas de la musique, la hauteur peut être perçue sur environ 6 octaves.

Les descripteurs utilisés sont comparés à des valeurs seuils afin de pouvoir décider de leur classe d'appartenance.

Les auteurs s'intéressent maintenant aux segments qu'ils viennent de détecter comme parole. Ils proposent une méthode de segmentation en locuteurs à l'aide d'une méthode existante [RR95] qu'ils ont adaptée. Cette méthode est basée sur l'utilisation de modèles de mélange de gaussiennes (GMMs) et des coefficients MFC comme descripteurs. Un modèle GMM est créé pour chaque locuteur que l'on souhaite caractériser. La méthode proposée offre de bonnes performances pour une durée de parole au moins égale à 5-10 secondes.

Enfin, les auteurs s'intéressent aussi à la segmentation par genre selon que ce soit une femme ou un homme qui parle. La méthode mise en œuvre pour résoudre ce problème s'inspire fortement des travaux de [PC96]. À la différence près qu'ici les auteurs utilisent à nouveau des GMMs au lieu des modèles de Markov cachés (HMM). L'algorithme consiste à caractériser la hauteur de la parole et prendre la décision sur le genre sachant que pour un homme la voix est plus grave donc contenue dans une bande de fréquences équivalente à 60-120Hz alors que pour les femmes la bande de fréquences est de l'ordre de 120-200Hz.

Les expérimentations ont été menées sur différentes bandes sonores. Les auteurs ont effectué des prises de son de personnes qu'ils connaissent pour l'identification et la reconnaissance du genre. Une vidéo a aussi été enregistrée avec quatre personnes : deux hommes et deux femmes. Les tableaux des figures 17, 18, 19 et 20 suivants résument les résultats des différents algorithmes développés au cours de cet article. Les performances affichées sont très satisfai-

Nom du locuteur	Nbr de segments analysés	Nbr de segments bien détectés
Mohanty	15	9
Nisha	11	11
Lalitha	16	10
Prakash	9	7

Fig. 17 – Résultats de la reconnaissance des locuteurs avec 5 secondes de temps d'identification et un GMM d'ordre 16 [VRSB99]

santes, toutefois la taille de la base de test est petite. Des tests à plus grande échelle doivent être menés afin de valider ces méthodes.

Nom du locuteur	Nbr de segments analysés	Nbr de segments bien détectés
Mohanty	15	9
Nisha	11	11
Lalitha	16	10
Prakash	9	7
Sarat	13	13

FIG. 18 – Résultats de la reconnaissance des locuteurs avec 10 secondes de temps d'identification et un GMM d'ordre 32 [VRSB99]

Nom du locuteur	Nbr de segments analysés	Nbr de segments bien détectés
Mohanty	122	115
Nisha	130	109
Lalitha	128	92
Prakash	81	75
Sarat	130	122

FIG. 19 – Résultats de la détection des segments de parole [VRSB99]

Nom du locuteur	Nbr de segments analysés	Nbr de segments bien détectés
Mohanty (H)	15	11
Nisha (F)	17	14
Lalitha (F)	17	13
Prakash (M)	16	15

FIG. 20 – Résultats de la reconnaissance de genre du locuteurs [VRSB99]

Harb et Chen [HS02]

Dans cet article, les auteurs proposent une nouvelle approche de description des scènes dans une vidéo en utilisant *exclusivement le flux audio*. Ainsi, les auteurs proposent de donner à chaque scène une signature basée sur une classification du son en parole/musique/bruit/silence. Cette méthode s'apparente sur beaucoup de points à la méthode précédente de Venugopal et al. [VRSB99] si ce n'est que Venugopal ne considérait pas la classe bruit dans sa classification.

La classification audio repose sur l'extraction d'un ensemble d'informations de haut niveau sémantique à partir de descripteurs bas niveau. Les descripteurs bas niveau utilisés sont :

- le Silence Crossing Ratio (SCR) : Le nombre de fois où il y a la présence d'un silence dans une fenêtre d'une durée de deux secondes.
- le Suivi de Fréquence (FT) : Le but de ce descripteur est de mesurer l'harmonicité d'un signal, aussi dans une fenêtre d'analyse d'une durée de deux secondes.
- la Distance entre Images Audio (DBAI) : Il s'agit de découper des parties du spectre d'une taille de 40×40 pixels donnant ainsi des images. Après, les auteurs calculent une distance entre ces images afin de caractériser d'éventuels changements dans le spectre. De plus, les auteurs découpent ces images en blocs de 8×8 pixels et calculent cette même distance bloc à bloc pour deux images consécutives. La distance finale entre deux images est la somme des distances bloc à bloc. Ce descripteur est extrait toutes les deux

secondes également.

– les Variations du Centroïde Spectral (FCV) : Le centroïde spectral est le centre de gravité du spectre et l'on étudie ses variations afin de mettre en valeur la dynamique du spectre. Plus précisément, les auteurs considèrent la moyenne des dérivées du centroïde spectral sur une durée de deux secondes.

La méthode de classification utilisée est un réseau de neurones. Le processus utilisé est le suivant. Les auteurs soumettent la vidéo à un détecteur de scènes [MAC01] et récupèrent les frontières de début et de fin pour chaque scène détectée. Les scènes sont découpées en segments d'une durée de deux secondes, et pour chaque segment l'ensemble des descripteurs énoncés ci-dessus est calculé. Ce vecteur de descripteurs est soumis au réseau de neurones qui retourne la probabilité d'appartenance à chacune des classes Parole, Musique, Bruit et Silence. La classe retenue est celle qui récolte le plus haut pourcentage.

Les expérimentations se sont portées sur la classification selon trois classes : Parole, Musique et Bruit. Le réseau de neurones a été entraîné sur un corpus de 30 minutes composé de 10 minutes de parole, 10 minutes de musique et 10 minutes de bruit. Le tableau de la figure 21 résume les performances de la classification dans ce cas donné. Les auteurs ne nous renseignent

Classe audio	Temps en minutes	Temps classifié en minutes	Rappel
Musique	6	5	83%
Parole	12	10.4	87%
Bruit	12	11.2	93%

Fig. 21 – Résultats de la classification en Parole, Musique et Bruit [HS02]

pas sur le taux de fausses alarmes de leur méthode de classification. En revanche les valeurs de rappel sont assez convenables bien que la taille du corpus de test soit assez petite. Toutefois, le corpus de test ainsi que celui d'apprentissage est assez varié et représentatif. Des tests à plus grande échelle seraient tout de même appréciables.

3.2.3 Indexation audio : moteur d'indexation

Dans cette partie, nous présentons une application directe de l'indexation. Il s'agit de la mise en place d'un moteur d'indexation dans le but de gérer une base de données audio et de pouvoir faire des requêtes à cette base de données sous différentes formes. De nombreux travaux, sur ce thème, sont présents dans la littérature comme par exemple [HC03a, Pie02, Pau02, SBY04]. Comme les méthodes utilisées dans ces études sont très similaires, nous détaillons les travaux de [HC03a] à titre d'exemple.

Dans les travaux de [HC03a], les auteurs proposent un moteur d'indexation de la bande sonore de programmes audiovisuels qu'ils ont appelé CYNDI. La figure 22 illustre l'architecture de CYNDI. Comme le montre la figure 22, CYNDI est composé de :

1. Un module de démultiplexage qui sépare le signal sonore d'un programme audiovisuel à l'entrée.
2. Un module de segmentation en parole ou musique en temps réel avec un délai de 15 secondes.
3. Un module d'indexation/segmentation de la parole.

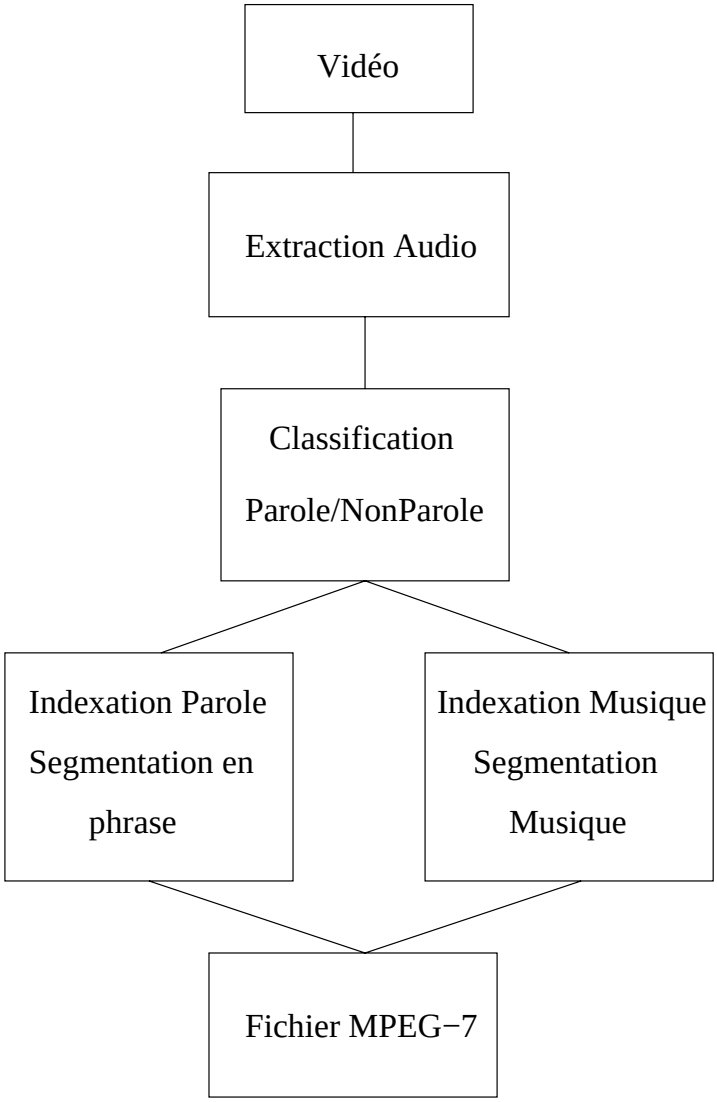


Fig. 22 – Architecture générale du moteur CYNDI [HC03a]

4. Un module d'indexation par le contenu des segments de musique.
5. Un module de génération des fichiers MPEG-7.

En ce qui concerne la classification entre parole et non parole, les auteurs utilisent la méthode de classification du signal audio présentée dans [HC02]. Ainsi les auteurs récupèrent les segments parole afin de les indexer. Les segments de non parole sont ensuite sous-classifiés en musique et non musique afin de faire une indexation de la musique tout comme pour la parole.

L'indexation de la parole se décompose en deux phases : la segmentation en paragraphes par seuillage dynamique et la génération des mots clés. La première phase est basée sur l'extraction et l'analyse de descripteurs bas niveau à partir du flux non compressé. Plus précisément, les auteurs proposent dans un premier temps d'utiliser une mesure de similarité entre deux fenêtres d'analyse du signal. Lorsque cette mesure dépasse un certain seuil alors

une frontière est détectée. Cette approche a déjà été développée, entre autre, dans [SJBS97]. Mais cette solution a pour inconvénient d'être dépendante du contenu. Les auteurs étudient aussi une solution statistique basée sur le critère d'information bayésien (BIC) utilisée dans les travaux de [GLA01]. Cette méthode consiste à rechercher les points du signal qui sont le plus probablement des points de changement de caractéristique acoustique, tel que le changement de locuteur. Finalement, les auteurs utiliseront la méthode basée sur la mesure de similarité en utilisant la distance de KullBack-Leibler (voir Annexe B.1) avec une méthode de calcul dynamique du seuil.
La méthode de calcul du seuil dynamique est basée sur l'étude de la distribution des longueurs de phrase. Les longueurs des segments de phrase correspondent aux changements de locuteur et de canal ainsi que le changement de phrase classique pour le même locuteur.

La génération des mots clés est obtenue à l'aide d'un moteur de Reconnaissance Automatique de la Parole (RAP) sur les segments de phrase obtenus précédemment. Plus l'occurrence d'un mot est importante lors d'un dialogue plus il est reconnu par le RAP. Ainsi, pour qu'un mot soit bien détecté il faut que le mot soit répété plus de 2 fois et qu'il contienne plus de 2 syllabes.

La dernière phase de l'indexation audio concerne l'indexation de la musique. Cette indexation est réalisée dans le but de gérer une base de données afin de pouvoir effectuer des requêtes de similarité sur le type de musique. Dans le moteur CYNDI, les auteurs utilisent l'approche basée sur la distance KullBack-Leibler (KL), au même titre que la parole, pour l'indexation des segments non parole. La distance KL est utilisée sur des fenêtres d'analyse d'une seconde ce qui permet la création d'index pour ces segments en se basant sur les vecteurs de moyenne et de variance du spectre dans les fenêtres d'une seconde.

Enfin, les auteurs proposent une implémentation de ce moteur qui permet une analyse et une indexation en temps réel, pour une durée de quatre heures par jour, des programmes d'une chaîne télévisuelle d'informations. La base de données ainsi constituée contient dès lors plus de 400 heures de programmes indexés par mots clés avec possibilité de navigation intelligente parmi les paragraphes d'un document.

3.2.4 Autres domaines d'application de l'indexation audio

Dans cette dernière partie, nous présentons quelques unes des autres applications de l'indexation audio. Parmi la diversité des travaux, dont notamment [AO98, PSAO02b], présents dans la littérature, nous avons sélectionné les travaux de [TC04, Ero01, HC03b].

Dans [TC04], l'auteur décrit le développement d'un algorithme de classification audio, plus particulièrement de la reconnaissance du genre du locuteur, pour les bandes sonores de journaux télévisés dans le cadre de la campagne d'évaluation TREC 2003. Toutefois, les auteurs se focalisent plus sur le processus de construction de ce classifieur que sur le résultat en lui-même. La première étape consiste à extraire les descripteurs audio avec des fenêtres d'une durée de 20 millisecondes. Les descripteurs bas niveau utilisés sont le centroïde spectral, les coefficients MFC, le RollOff ainsi que les Coefficients de Prédiction Linéaire (LPC) [Mak95]. Concernant la classification, les auteurs ont utilisé et adapté une méthode basée sur l'utilisation des supports vecteurs machine (SVM) [DHS00]. Cette méthode de classification par SVM est une méthode de classification supervisée.

Les expérimentations ont été menées sur le corpus fourni par TRECVID 2003, ce qui correspond à 120 heures de vidéo de journaux télévisés, les résultats sont de l'ordre de 78% de rappel pour la détection des voix féminines et 74% de rappel pour les voix masculines. Ces résultats sont très corrects compte tenu de la taille et de la richesse du corpus de TRECVID. Les auteurs précisent que des résultats plus précis sont présents dans [Ha03].

Dans la thèse de [Ero01], l'auteur propose notamment une nouvelle application de l'indexation audio. En l'occurrence, la reconnaissance automatique des instruments de musique. Le schéma de la figure 23 illustre le mode de fonctionnement du système mis en place. Parmi

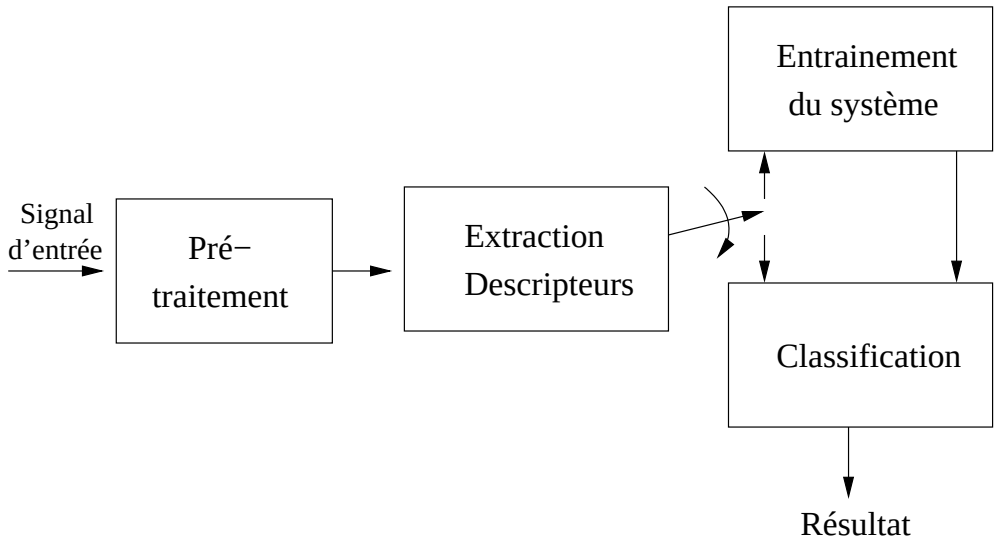


FIG. 23 – Schéma de fonctionnement du système de reconnaissance des instruments de musique [Ero01]

les descripteurs utilisés, il y a les coefficients MFC, le centroïde spectral et un algorithme de détection de la fondamentale développé par [Tal95].
Dans ses travaux de thèse l'auteur a exploré différentes méthodes de classification telles que la méthode des K-voisins les plus proches (KNN), ou bien les modèles de mélange gaussien (GMM). L'auteur a aussi utilisé le critère de Fisher [Mar98] pour réduire de manière optimale l'ensemble des descripteurs à soumettre aux méthodes de classification.
Les expérimentations ont été menées sur une base de données musicale qui comprenait 1498 instruments à reconnaître. Le tableau de la figure 24 résume les résultats globaux obtenus. L'auteur nous donne les performances uniquement concernant le rappel mais ne nous renseigne

Méthode de classification	Rappel
KNN	77%
GMM (d'ordre 4)	69.7%

FIG. 24 – Résultats généraux de la reconnaissance d'instruments

pas sur la précision de sa méthode de reconnaissance à savoir les fausses alarmes. Compte tenu d'une base de test suffisamment riche et de taille convenable, nous pouvons conclure sur la robustesse du système présenté ici.

Enfin, dans [HC03b] les auteurs nous présentent une méthode de reconnaissance du genre

des locuteurs dans les segments de parole des bandes audio. La méthode proposée ici consiste à classifier les segments parole en deux classes : locuteur homme et femme.
Les descripteurs utilisés ici sont évidement propres à l'étude de la parole. Il y a donc les coefficients MFC extraits sur des fenêtres d'analyse de 10ms. Les auteurs utilisent un réseau de neurones à plusieurs couches comme méthode de classification.
Les expérimentations ont été effectuées sur une base de données constituée de plusieurs enregistrements effectués sur des stations de radio françaises et britanniques. L'ensemble de la base de données de test a une durée totale de 120 minutes, soit 2 heures. Le tableau de la figure 25 résume l'ensemble des résultats obtenus. Les performances sont équivalentes qu'il s'agisse

Source	Rappel Classe Homme	Rappel Classe Femme
Radio française	94.70%	88.75%
Radio britannique	93.14%	89.14%

Fig. 25 – Résultats globaux de la reconnaissance de genre du locuteur

de la langue française ou de la langue britannique. Il est possible de dire que le système ne dépend pas du langage parlé. De plus, les performances obtenues sont d'assez bonne facture au regard de la base utilisée de par sa taille et sa richesse.

3.3 Conclusion

En conclusion, dans ce chapitre, nous avons vu divers descripteurs audio. Un certain ensemble de descripteurs est commun à plusieurs travaux et permet une classification complète en Parole, Musique, Bruit et Silence. D'autres descripteurs, comme le kurtosis, sont plus spécifiques pour certaines tâches. Le choix des descripteurs est conditionné par les objectifs de l'indexation de la bande sonore. Nous allons voir dans la suite de cette thèse le choix des descripteurs à faire pour notre tâche d'indexation cross-média.

Chapitre 4

État de l'art des méthodes d'indexation cross-média

Dans ce chapitre, nous nous intéressons aux méthodes présentes dans la littérature traitant de l'indexation des contenus audiovisuels par l'analyse cross-média : analyse conjointe des flux audio et vidéo. Cette dernière constitue un thème de recherche émergeant dans le vaste domaine de l'indexation des contenus multimédia artistiques. Nous commençons cette étude par les premières tentatives de fusion des flux audio et vidéo [SL97], puis d'autres approches cross-média d'indexation vidéo qui requièrent des algorithmes d'analyse et d'indexation audio plus approfondies sont présentées par la suite [TP99, HClK+03].

Saraceno et Leonardi [SL97]

Dans [SL97], les auteurs proposent une méthode d'indexation vidéo en scènes basée sur l'analyse conjointe audio et vidéo. Ces travaux sont parmi les premiers dans le domaine de l'indexation vidéo en scènes par l'analyse cross-média.

Le principe global de la méthode est illustré par la figure 26 : VSCD pour Video Shot Cut Detector, SCD pour Shot Cut Detector, VFE pour Video Features Extractor, SC pour Scene detector and Characterizator, SD pour Silence Detector, SPD pour Speech Detector et MD pour Music Detector : Comme illustré par la figure 26, les flux audio et vidéo sont traités de manière parallèle et indépendante. Concernant le flux audio, les auteurs proposent une classification en quatre classes : Silence, Parole, Musique et Bruit. Concernant la classification, la méthode proposée est multi-passes. Tout d'abord, les segments de silence sont détectés au regard du niveau d'énergie. Lors de la seconde passe, les auteurs analysent la mesure d'autocorrélation pour déterminer les segments de parole parmi ceux qui n'ont précédemment pas été étiquetés comme silence. Après quoi, lors de la troisième passe les segments musicaux sont caractérisés à l'aide de la mesure d'autocorrélation. Enfin les segments restant sont automatiquement marqués comme des segments de bruit. Concernant l'analyse et l'indexation du flux vidéo, il s'agit d'estimer la différence des valeurs de luminance entre deux images successives afin de pouvoir caractériser les frontières des plans de montage.

La prochaine étape de la méthode présentée dans cet article concerne la fusion des informations issues des analyses audio et vidéo. Ainsi, comme le montre la figure 26, le bloc

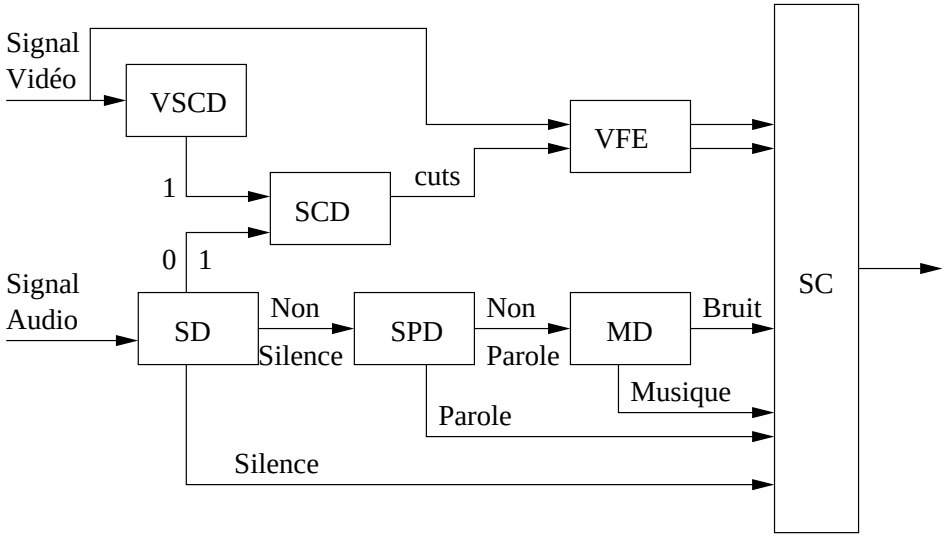


FIG. 26 – Schéma général de la détection des scènes

VSCD assigne deux valeurs de probabilité pour chacune des images de la vidéo. Ces deux valeurs correspondent à la probabilité d'être une transition abrupte et la probabilité d'être une transition graduelle. Ces valeurs probabilistes sont affectées en fonction de la valeur de la différence de luminance entre deux images successives. Dans le cas d'un fort pic de la fonction des valeurs de luminance, la probabilité d'une transition abrupte affectée à l'image courante sera forte. Pour ce qui est de l'attribution de la probabilité d'une image appartenant à une transition graduelle, le principe reste le même. Ensuite, ce résultat est envoyé au bloc SCD ainsi qu'un booléen venant de SD selon que l'image courante est associée ou non à un segment de silence. Ainsi, à l'aide de deux seuils, λ_1 pour les transitions abruptes et λ_2 pour les transitions graduelles, les auteurs caractérisent et classifient les transitions vidéo. Puis le résultat de la détection des transitions abruptes est envoyé au module VFE (voir figure 26) qui va extraire des descripteurs vidéo par l'identification de l'objet le plus important pour chaque plan de montage. La caractérisation de l'objet est obtenue à partir de la combinaison d'un algorithme de segmentation et de suivi. Le résultat est envoyé au module SC ainsi que le résultat de la classification du flux audio.

Enfin, la dernière phase de la méthode est la fusion des descripteurs audio et vidéo. Le principe de fusion, géré par le module SC, consiste à estimer une valeur de corrélation entre les plans de montage adjacents. Une fois que toutes les scènes ont été créées, le module de fusion extrait une image clé de chacune des scènes ainsi que d'autres informations permettant de les caractériser. Ces informations sont le nombre d'objets présents, le nombre de locuteurs, le type de musique, etc ...

Les expérimentations menées dans le cadre de ces travaux ont été effectuées sur une base de données vidéo composée de :

- 5 minutes de journaux télévisés avec des reportages courts et des transitions abruptes ;
- 5 minutes de vidéo composées de 7 publicités toutes différentes avec des transitions vidéo abruptes et graduelles et
- 10 minutes de film comportant des transitions vidéo abruptes et graduelles.

Les auteurs ne nous renseignent pas sur les valeurs de rappel et précision. Ne sont mentionnés que le pourcentage de transitions abruptes correspondantes avec un silence et une estimation des performances de la méthode globale en fonction du type de genres vidéo. Ainsi, les auteurs concluent sur le fait que l'ajout de l'analyse du flux audio est un apport important pour obtenir de bonnes performances de détection. De plus, les auteurs précisent que le détecteur est plus robuste pour les journaux télévisés et les publicités que pour les films.

Alatan, Akansu et Wolf [AAW01]

Dans [AAW01], les auteurs proposent d'étudier un modèle spécifique de scène : les scènes de dialogue. La méthode proposée fait appel à la fois à l'analyse du flux vidéo, détection des frontières de plan, détection de visage, à l'analyse du flux audio et à la classification audio. Les modèles de Markov cachés sont utilisés pour la détection des frontières de scènes de dialogue. Les auteurs commencent par définir leur modèle de scène de dialogue comme étant une succession de plans de montage qui contiennent des conversations entre individus. Ainsi, les auteurs présentent une méthode qui, dans premier temps, détecte les frontières de plans de montage. Puis, en parallèle, la méthode tente de détecter les visages humains présents dans la scène ainsi que la présence de parole dans le flux audio. L'approche doit donc traiter trois sources d'informations : la présence de personnes, la présence de parole dans le flux audio et le lieu où se déroule la scène. Ainsi, les auteurs proposent 3 niveaux de complexité de détection : bas, moyen et haut. Plus le niveau de complexité est élevé meilleure est la détection des scènes car plus d'informations provenant de l'analyse des flux multimédia sont extraites. Dans le cadre de cet article, les auteurs vont considérer le niveau moyen de complexité nécessitant une classification audio en parole, musique et silence pour la partie son, et une détection des plans et des visages pour la partie vidéo. Pour mener à bien cette tâche, les auteurs utilisent les modèles de Markov cachés. Les auteurs proposent deux topologies différentes pour modéliser les scènes de dialogues (voir figure 27). La nomenclature dans la figure 27 est la suivante : EST pour *établissement d'une scène*, DIA pour *scène de dialogue*, TRA pour *scène de transition* et NON *pour Non-scène de dialogue.*

Concernant le modèle gauche-à-droite, les auteurs justifient leur choix en disant que les films peuvent être modélisés comme une succession de scènes de dialogue et de transition (pas de dialogue) avec un état appelé établissement de scène au tout début. Le modèle circulaire est plus simple puisqu'il décrit un film comme étant composé de scènes de dialogue et de scènes sans dialogue. L'inconvénient du premier modèle c'est qu'il faut connaître le nombre de scènes du contenu a priori. Les résultats qui sont présentés proposent une comparaison des performances de détection des scènes de dialogue en fonction de la topologie du modèle de Markov caché. Dans tous les cas, les tests ont été effectués sur un corpus vidéo MPEG-7 dont les auteurs ne nous renseignent pas sur sa taille. Pour ce qui est de la comparaison des performances entre les deux topologies des modèles de Markov elles offrent toutes les deux un niveau de performance équivalent. D'autant plus que ce niveau de performance est assez élevé. Les travaux présentés ici reposent sur un modèle particulier de scène, par conséquent, cette approche ne permet pas de détecter toutes les scènes d'un contenu multimédia. De plus, pour l'instant la méthode proposée n'est pas capable de différencier entre les scènes de monologue et celles de dialogue.

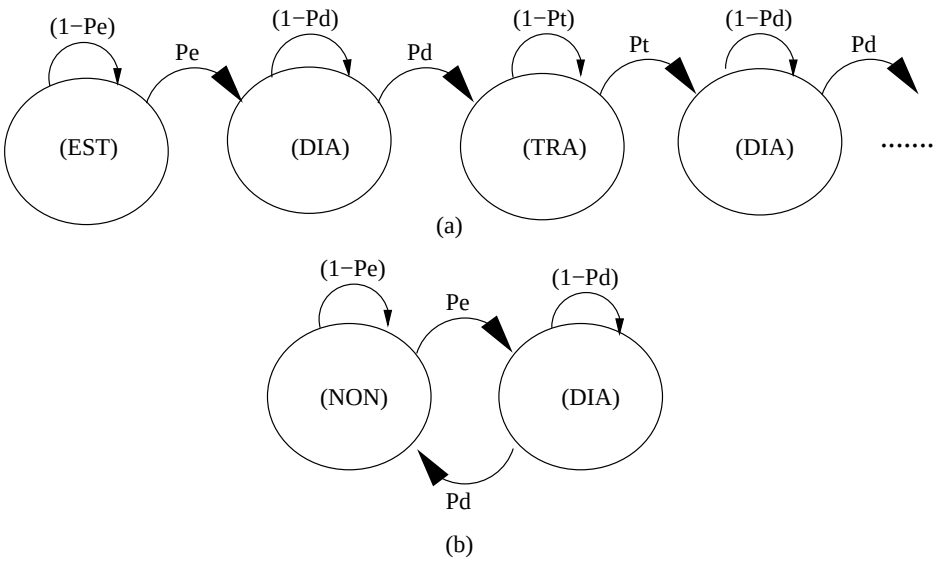


FIG. 27 – Diagrammes d'état des HMM pour la modélisation des scènes de dialogues dans les films : Modèles (a) Gauche-à-droite et (b) Circulaire

Hsu, Chang, Huang, Kennedy, Lin et Iyengar [HClK+03]

Dans [HClK+03], les auteurs présentent une méthode cross-média pour la découpage de journaux télévisés en scènes. Dans cet article, les auteurs définissent deux types de scène :

- Les scènes qui correspondent aux segments temporels vidéo traitant du journal télévisé en lui-même et
- Les scènes qui sont composées du reste comme les publicités, les interludes, etc ...

Le but de ces travaux est donc de retrouver et de classifier ces deux types de scènes. Pour cela, l'approche proposée est basée sur un modèle statistique permettant la fusion des données afin d'estimer un évènement a posteriori en l'occurrence une frontière de scène. Ce modèle utilise des données extraites d'une base d'apprentissage et fusionne ces données pour prendre une décision sur la présence ou non d'une frontière de scène. Le modèle estimé est représenté par $q_\lambda(b|x)$, où $b \in \{0,1\}$ est une variable aléatoire correspondant à la présence ou non d'une frontière de scène dans le contexte x et λ l'ensemble des paramètres estimés. Dans ces travaux, x représente une image pouvant être labellisée comme étant une limite de scène. Pour chaque x, un ensemble de descripteurs binaires, $f_i(x,b) = 1_{\{g_i(x)=b\}} \in \{0,1\}$, est calculé. Où $1_{\{.\}}$ est une fonction d'indication et g_i est une valeur prédicative d'une frontière de scène utilisant le i-ème descripteur. Ce modèle s'appelle le modèle de l'entropie maximale. Nous pouvons ainsi résumer ce modèle par l'équation suivante :

$$q_\lambda(b|x) = \frac{1}{Z_\lambda(x)} exp^{\sum_i \lambda_i f_i(x,b)} \tag{76}$$

avec $\sum_i \lambda_i f_i(x,b)$ une combinaison linéaire des descripteurs binaires avec des paramètres réels λ_i qui permettent de pondérer l'importance du descripteur d'index i et $Z_\lambda(x)$ un vecteur de normalisation.

L'étape suivante consiste à estimer les paramètres de l'équation 76. Pour cela, les auteurs utilisent une méthode qui consiste à minimiser la valeur de la distance de Kullback-Leibler (voir Annexe B.1) calculée à partir de la base d'apprentissage qui suit une distribution statistique empirique P. Ce qui donne l'équation suivante :

$$\lambda^* = argmin_\lambda D(P||q_\lambda) \tag{77}$$

avec $D(.||.)$ la mesure de la distance de Kullback-Leibler définie par :

$$D(\tilde{p}||q_\lambda) = \sum_x P(x) \sum_{b\in\{0,1\}} P(b|x) log \frac{P(b|x)}{q_\lambda(b|x)} \tag{78}$$

Le problème posé par l'équation est équivalent à maximiser la vraisemblance de le domaine logarithmique :

$$L_P(q_\lambda) = \sum_x \sum_b P(x,b) log q_\lambda(b|x) \tag{79}$$

La méthode d'estimation des paramètres λ_i est itérative. Le processus d'estimation itère jusqu'à la minimisation de la divergence (voir équation 78). À chaque itération, nous avons $\lambda_i^{'} = \lambda_i + \Delta\lambda_i$ avec :

$$\Delta\lambda_i = \frac{1}{M} log \left\{ \frac{\sum_{x,b} P(x,b) f_i(x,b)}{\sum_{x,b} P(x) q_\lambda(b|x) f_i(x,b)} \right\} \tag{80}$$

où M est une constante qui permet de contrôler la rapidité de la convergence.

Après quoi, les auteurs proposent une méthode de sélection des descripteurs audio et vidéo les plus pertinents parmi un large ensemble de descripteurs. Pour cela, étant donné un ensemble de descripteurs candidats C et un modèle initial d'entropie maximale q, ce modèle peut-être enrichi par l'ajout d'un nouveau descripteur $h \in C$ et de sa pondération associée α. Ce qui donne :

$$q_{\alpha,h}(b|x) = \frac{exp^{\{\alpha h(x,b)\}} q(b|x)}{Z_\alpha(x)} \tag{81}$$

avec $Z_\alpha(x)$ un vecteur de normalisation.
Le descripteur sélectionné h^* à chaque itération est défini par :

$$\begin{aligned} h^* &= argmax_{h\in C}\{sup_\alpha\{D(P||q) - D(P||q_{\alpha,h})\}\} \\ &= argmax_{h\in C}\{sup_\alpha\{L_P(q_{\alpha,h}) - L_P(q)\}\} \end{aligned} \tag{82}$$

Parmi ces descripteurs candidats, il y a notamment :

- la fréquence d'apparition du présentateur de télévision,
- pour chaque image, une booléen pour indiquer si nous sommes dans une partie publicitaire ou bien dans le programme de télévision à proprement parlé,
- les variations des valeurs pour la hauteur dans les zones voisées permettant de déterminer, suivant l'intonation, le début ou la fin des phrases par exemple.
- la caractérisation des pauses dans la bande audio et
- la durée des segments de paroles ainsi que le débit de parole du locuteur.

Enfin, les tests présentés ont été menés à grande échelle dans le cadre de la campagne d'évaluation TRECVID 2003. La durée totale de la base de données vidéo est de 111 heures qui ont été réparties de la manière suivante : 66 heures pour l'entraînement du système et les 45 heures restantes pour les tests. Les performances de détection et de la classification des segments appartenant aux journaux et ceux qui n'y appartiennent pas sont de l'ordre de 73% de rappel et de 80% de précision. Ces résultats sont d'autant plus probants que la quantité de données de la base de test est grande.

Sundaram et Chang [SC00]

Les auteurs de [SC00] exposent une approche de détection des scènes en réalisant une fusion des résultats de deux segmentations : audio et vidéo. Ils commencent tout d'abord par présenter leur modèle de scène audio et vidéo. Une scène audio est modélisée comme une collection de quelques sources sonores dominantes. Ces mêmes sources sont supposées constantes tout au long d'une scène et peuvent être caractérisées par quelques descripteurs. Un changement de scène audio est caractérisé par une variation des valeurs des descripteurs. De même, une scène vidéo est modélisée par un ensemble de plans ayant un même lien sémantique. Les auteurs considèrent la similitude de la chrominance comme critère de cohérence des plans. Évidemment, un changement de scène correspond à une variation dans les valeurs de chrominance. Enfin, les auteurs considèrent une scène comme étant la corrélation temporelle d'une scène audio et vidéo. Les frontières audio et vidéo qui ne sont pas en correspondance temporelle sont appelées *singleton*. Les auteurs exposent un modèle à mémoire, de type FIFO (First-In-First-Out), d'un suiveur composé de deux paramètres :

- La mémoire : la quantité totale, T_m, d'informations contenues dans le suiveur et
- Le champ d'attention (attention span) : un petit espace de temps, T_{as}, qui correspond aux données récentes contenues dans le suiveur.

Le principe réside sur le fait que le suiveur compare les données récentes (T_{as}) avec la globalité de ses données (T_m) pour décider de la présence ou non d'une frontière de scène.

Concernant la détection des frontières de scène audio, les auteurs présentent un algorithme basé sur l'extraction et l'analyse de descripteurs audio. Ainsi, une frontière de scène audio est déterminée par le seuillage d'une fonction de décision, D, définie pour chaque instant de temps par :

$$D(t_0) = \sum_i b_i \tag{83}$$

avec b_i un paramètre d'affaiblissement exponentiel du descripteur audio i.
Les paramètres b_i sont déterminés par l'égalité suivante :

$$C_i(t) = exp^{b_i t} \tag{84}$$

avec C_i la fonction de corrélation pour le descripteur i.
Cette fonction de corrélation est définie pour chacun des descripteurs, f, par :

$$C_f(m\delta) = 1 - d(f(t_0, t_0 - t_{as}), f(t_0 + m\delta, t_0 + m\delta - t_{as})) \tag{85}$$

avec $f(t_1, t_2)$ l'enveloppe qui correspond au descripteur f pour la durée $[t_1, t_2]$, $m \in [-N...0]$ où $N \equiv (T_m - T_{as})/\delta$, δ le décalage de la fenêtre d'analyse et d la métrique euclidienne sur

les enveloppes. La métrique utilisée est intuitive puisqu'il s'agit d'une comparaison point par point entre les deux enveloppes.
Comme la méthode nécessite un calcul de corrélation entre les descripteurs, les auteurs définissent l'enveloppe d'un descripteur comme étant une modélisation de la courbe de ses valeurs. Pour déterminer cette enveloppe les auteurs procèdent à différents tests suivant que la courbe des valeurs du descripteur soit de type linéaire, constante, quadratique, exponentielle, hyperbolique ou bien une somme d'exponentielles. Pour déterminer la meilleure approximation les auteurs retiennent celle qui minimise l'erreur des moindres médians. À noter que cette méthode d'enveloppe n'est appliquée qu'aux descripteurs ayant des valeurs scalaires. Les autres étant utilisées dans leur formes brutes.

À propos de la détection des scènes vidéo, les auteurs ont développé deux notions : *rappel* et *cohérence*. Le principe de rappel entre deux plans de montage a et b est formalisé par :

$$R(a,b) = (1 - d(a,b)) \times f_a \times f_b \times (1 - \Delta t / T_m) \tag{86}$$

avec $R(a,b)$ la mesure de rappel entre les plans a et b, $d(a,b)$ une distance d'histogrammes de couleur entre les images clés correspondant aux deux plans a et b, f_i le ratio entre la longueur d'un plan i avec la taille T_m et δt la différence temporelle entre les deux plans.
Ainsi, une transition à l'instant t_0 est détectée si la valeur de rappel entre le plan qui précède et celui qui succède est inférieure à un seuil. De plus, les auteurs calculent une valeur de cohérence qui est définie par :

$$C(t_0) = \frac{\left(\sum_{\forall a \in T_{as}, b \in \{T_m \ T_{as}\}} R(a,b)\right)}{C_{max}(t_0)} \tag{87}$$

avec $C(t_o)$ la valeur de cohérence au niveau de la frontière t_0 qui correspond à la somme des valeurs de rappel entre toutes les paires de plans de montage au niveau de la borne t_0 et $C_{max}(t_0)$ est obtenu en fixant $d(a,b) = 0$ dans la formule 86.

Enfin, la méthode calcule la valeur de cohérence entre chaque paire adjacente d'images clés. Si une transition à l'instant t_0 détectée précédemment correspond à un minima local de C alors cette transition est étiquetée comme transition de scène vidéo. La dernière étape de la méthode globale consiste à regrouper les résultats des scènes audio et vidéo afin de caractériser les scènes générales. Pour cela, les auteurs utilisent un algorithme simple de recherche du plus proche voisin. Cet algorithme prend en compte les ensembles des frontières de scènes audio et vidéo et recherche les frontières audio et vidéo qui correspondent temporellement avec une tolérance sur le décalage temporel. Toutefois, les auteurs proposent de faire évoluer ce système avec l'utilisation d'une méthode statistique pour la prise de décision. Cela consiste à calculer tous les écarts entre chacune des frontières de scènes audio et vidéo, puis à étiqueter et à apprendre une partie de ces écarts dans le but de déterminer une distribution empirique afin de déterminer statiquement si un écart entre une frontière audio et vidéo correspond à une frontière de scène.

Les résultats expérimentaux présentés dans ce papier ont été menés sur la première heure d'un film de fiction. Les meilleures performances ont été obtenues avec $T_m = 31$ secondes et $T_{as} = 16$ secondes dans le cas de la détection des frontières de scènes audio et $T_m = 32$ secondes et $T_{as} = 16$ secondes pour les scènes vidéo. Ainsi, en utilisant une tolérance temporelle de 4 images pour la recherche des frontières de scènes les auteurs obtiennent un rappel de 84%. Cependant, les auteurs ont comparé cette méthode avec l'utilisation d'une méthode de décision

statistique et ils montrent que le rappel est inférieur, de l'ordre de 75%. Mais les auteurs prétendent que ceci est dû au fait qu'ils travaillent sur un petit ensemble de scènes et qu'un ensemble de test plus conséquent améliorerait les performances.

Kijak, Gravier, Gros, Oisel et Bimbot [KGG+03]

Dans [KGG+03], il est présenté une méthode pour la caractérisation des frontières de scène des vidéos de tennis basée sur les modèles de Markov cachés.

La première étape de l'algorithme consiste à segmenter les flux audio et vidéo. Le flux vidéo est segmenté en plans de montage en considérant les transitions de type cut et de type dissolution (dissolve). Pour chaque plan, une image clé est extraite. Dans le cas du flux audio, les auteurs proposent une classification en *impacts de balle*, *applaudissements* et *paroles*. Enfin, ces résultats sont traités par le modèle de Markov caché et chaque plan est alors classifié de manière hiérarchique suivant : *first missed serve*, *rally*, *replay* et *break*. Les auteurs ont déterminé cette classification hiérarchique à partir d'une grammaire vidéo des matchs de tennis mise au point par des observations. Plus précisément, les auteurs partent du principe que les retransmissions des matchs de tennis sont effectuées avec un nombre limité de caméras qui restent fixes. De plus, les matchs de tennis sont composés d'un nombre fini de scènes types que l'on retrouve plusieurs fois. D'où le choix des auteurs concernant les scènes types (voir ci-dessus). Les auteurs précisent aussi que les matchs de tennis sont composés de trois types de vues : globale, moyenne, rapprochée et le public. À chaque type de vue correspond un type d'action défini plus haut. Et les auteurs définissent une syntaxe, un scénario, pour des matchs de tennis à partir des prises de vue et des actions de jeu.

Pour ce qui est du domaine vidéo, la but de la méthode consiste à détecter les prises de vue globales par l'extraction et l'analyse de descripteurs vidéo. Pour commencer, les auteurs proposent une définition des prises de vue globales. En effet, les prises de vue globales sont caractérisées par une couleur assez homogène du contenu (la couleur du court de tennis). De plus, lors des prises de vue globales, le mouvement de la caméra est faible voir nul. À partir de ces observations, les auteurs proposent deux types de descripteurs vidéo pour caractériser les vues globales :

- Un vecteur des couleurs dominantes F et sa cohérence spatiale C et
- L'activité, A, de mouvement de la caméra de la prise de vue.

Ils définissent aussi une mesure de similarité entre les images extraites des différentes vues globales :

$$v(K_1, K_2) = w_1|C_1 - C_2| + w_2 d(F_1, F_2) + w_3|A_1 - A_2| \tag{88}$$

avec w_1, w_2 et w_3 les pondérations et $d(F_i, F_j)$ la distance entre les vecteurs F_i et F_j des couleurs dominantes.
En ce qui concerne le domaine audio, les auteurs proposent une méthode statistique de classification du flux audio en parole, applaudissements, impacts de balle, bruit et musique. Cette méthode considère un mélange de gaussiennes pour chacune des classes considérées. Chaque plan de montage est alors associé à un vecteur de données binaire qui spécifie l'appartenance à chacune des classes décrites ci-dessus. Enfin, compte tenu des informations précédentes les auteurs définissent 4 entités de bases d'un match de tennis : 2 considérées comme faisant partie de la partie en elle-même (first missed serves et rally) et les 2 autres ne faisant pas partie du

match (breaks et replays). Chacune de ces 4 entités est modelée par un modèle de Markov caché et chaque état du modèle modélise une transition vidéo de type cut ou dissolution. Les auteurs définissent une observation comme étant composée de :

- Une mesure de similarité v_t entre les plans,
- La durée du plan d_t et
- Le vecteur audio a_t, défini ci-dessus.

De manière plus formelle, les auteurs définissent la probabilité d'une observation, o_t, sachant un état, j, par :

$$b_j(o_t) = p(v_t|j) \times p(d_t|j) \times P[a_t|j] \tag{89}$$

avec $p(v_t|j)$ et $p(d_t|j)$ donnés par un histogramme lissé et $P[a_t|j]$ le produit suivant chaque classe audio k de la probabilité discrète $P[a_t|j]$.
La segmentation et la classification de la séquence complète d'observations, $o = o_1 o_2 ... o_T$, en éléments structurels sont effectuées en même temps à l'aide de l'algorithme de Viterbi. La séquence d'états la plus probable, $s = s_1 ... s_T$ est obtenue par :

$$\widehat{s} = argmax_s \quad ln(p(s)) + \sum_t ln(b_{s_t}(o_t)) \tag{90}$$

Deux heures de vidéos préalablement annotées manuellement ont été utilisées. De manière classique, la première moitié de la base vidéo a été utilisée pour l'apprentissage du modèle et l'autre moitié pour les tests. Les résultats sont présentés séparément en fonction de la détection des 4 entités définies : First serve, rallies, replay et break. Les valeurs de rappel sont, respectivement, de 41%, 65%, 82% et 92%. Quant aux valeurs de précision elles sont égales, respectivement, à 71%, 82%, 86% et 68%. Les performances affichées sont très encourageantes, toutefois la base de données multimédia est assez pauvre, elle mériterait d'être enrichie par des vidéos supplémentaires de matchs de tennis sur des terrains différents (terre battue, gazon, etc...).

Chen, Shyu, Liao et Zhang [CSLZ02]

Une méthode de détection des scènes basée sur l'extraction de descripteurs audio et vidéo est proposée dans [CSLZ02]. La méthode présentée dans ce papier peut se décomposer en deux grandes parties :

- La détection des frontières des plans de montage par une méthode de segmentation non supervisée et
- L'extraction et l'analyse de descripteurs audio à l'intérieur des plans vidéo détectés au préalable.

Enfin la décision finale de détection des scènes est prise suivant l'étude de la distance entre les plans et des descripteurs audio extraits. L'algorithme de détection des transitions vidéo est organisé de la manière suivante :

- Étape 1 : La différence pixel à pixel est calculée pour tout couple d'images consécutives. Si cette différence est supérieure à un seuil alors une transition vidéo est détectée sinon l'algorithme passe à l'étape 2.

- Étape 2 : Le méthode procède maintenant à l'extraction des objets en mouvement pour toutes les paires d'images consécutives [CSZK01]. Il est obtenu la génération de cartes de segmentation pour chaque couple d'images consécutives. Il ne reste plus qu'à calculer une distance entre chacune des cartes de segmentation puis appliquer un seuil sur ces distances pour déterminer si l'image courante est considérée comme frontière de plan ou non. Si ce n'est pas la cas, l'algorithme passe à l'étape 3.
- Étape 3 : Dans cette dernière phase, la méthode fait appel à un algorithme de suivi d'objets entre deux images successives. Si l'aire totale formée les objets suivis est plus petite qu'un seuil alors l'image courante est étiquetée comme étant une frontière de plan de montage.

La caractérisation des transitions audio est basée sur l'extraction et l'analyse de descripteurs audio tels que le volume(V), l'énergie(P), l'énergie dans les sous-bandes (Sub-P), le taux de passage par zéro (ZCR) , le centroïde de la fréquence (FC), les sous-bandes de fréquence (FB), le flux spectral (SF), le flux cepstral (CF) et un descripteur qui indique le nombre de fenêtres d'analyse d'un ensemble de fenêtres qui ont une énergie inférieure à l'énergie moyenne de toutes les fenêtres de l'ensemble (LSTER). Les auteurs proposent le découpage en sous-bandes de fréquences suivantes (avec fs le fréquence d'échantillonnage) : 0 - 1/16 fs, 1/16 fs - 1/8 fs, 1/8 fs - 1/4 fs et 1/4 fs - 1/2 fs. De part leurs observations, les auteurs ont remarqué que la première et la troisième bande étaient plus représentatives que les autres c'est pourquoi ils ont sélectionné ces deux bandes pour calculer l'énergie dans les sous-bandes.

Pour l'analyse du flux audio, les auteurs découpent la bande sonore en fenêtres d'analyse puis regroupent ces mêmes fenêtres en plans vidéo suivant le découpage en plans de montage effectué précédemment. Pour chacune des fenêtres d'analyse les vecteurs de descripteurs audio sont calculés. Si le volume est inférieur à 0.003 et le taux de passage par zéro supérieur à 50 alors cette fenêtre est étiquetée comme silence. De plus, si plus de 70% des fenêtres sont étiquetées comme silence dans un même plan alors le plan tout entier est étiqueté comme silence. De même si un plan est composé de plusieurs fenêtres étiquetées comme silence dont la durée totale est supérieure ou égale à 0.33 seconde alors ce plan est considéré comme un plan silence. Les plans marqués silence sont ignorés pour la suite du processus d'analyse. Après quoi, les auteurs proposent un regroupement des descripteurs audio en trois grands groupes : Volume, Puissance et Spectral. Le tableau suivant rappelle le nom des trois groupes ainsi que leur modélisation :

Nom du groupe	Modélisation
Volume	$Vec(shot_i) = mean(V_i) + dev(V_i) + vdr(V_i) + diff(V_i)$
Puissance	$Pvec(shot_i) = dev(P_i) + dev(Sub - P_i) + lster(P_i)$ $+lster(Sub - P_i) + diff(P_i) + diff(Sub - P_i)$
Spectral	$dev(FB_u) + dev(FC_i) + diff(SF_i) + diff(CF_i)$

FIG. 28 – Modèles des groupes de descripteurs audio

Dans la table précédente, *mean* représente la valeur moyenne d'un descripteur dans le plan numéro i, *dev* l'écart-type d'un descripteur dans le plan numéro i, *vdr* le volume dynamique dans le plan numéro i et $diff$ l'écart-type de la différence fenêtre-à-fenêtre d'un descripteur dans le plan numéro i. Grâce à la modélisation proposée par le tableau (figure 28), les auteurs définissent une distance entre deux plans voisins afin de caractériser les scènes. Cette distance

est définie par :

$$T_{sv} = \frac{(mean(dist(sv_i, sv_{i+1})) + dev(dist(sv_i, sv_{i+1})))}{\sqrt{(2)}} \tag{91}$$

où sv peut-être soit Vec, soit $Pvec$ ou $Svec$, sv_i et sv_{i+1} sont les valeurs de sv des deux plans voisins i et $i+1$ et $dist$ représente la distance Euclidienne.
Si cette distance est supérieure à un seuil alors la frontière de plan considérée est maintenant marquée comme étant une frontière de scène.

Les tests ont été menés sur une longue vidéo de film et des journaux télévisés. L'ensemble de ces vidéos comprend 191 scènes. Les résultats obtenus dans ces travaux sont de l'ordre 89% de rappel et 92% de précision. Ces résultats sont tout à fait convenables d'autant qu'ils ont été obtenus sur une taille de corpus suffisamment représentative. Cependant, les auteurs ne considèrent qu'un certain nombre de genres vidéo (films et journaux télévisés) et ils n'ont pas testé leur méthode sur des extraits de clip de musique ou bien sur des séries télévisées.

Leonardi et Migliorati [LM00]

Dans [LM00], les auteurs proposent deux approches pour la segmentation des documents audiovisuels. La première est basée sur l'utilisation d'un automate d'états finis utilisant les descripteurs bas-niveau extraits du flux compressé MPEG relatif au mouvement. La seconde méthode fait appel aux modèles de Markov cachés. Les travaux présentés dans ce papier ont été validés sur des contenus sportifs, plus particulièrement les matchs de football. Le but de ces recherches est d'extraire de manière automatique les événements importants qui sont présents dans les matchs de football comme : les coups francs, les buts, les coups de pied de coin (corner), etc...

Pour la première méthode par automates finis, il faut commencer par extraire les descripteurs bas-niveau tels que les vecteurs de mouvement directement extraits du flux compressé MPEG-2. Puis ces descripteurs sont alors utilisés par les algorithmes de prise de décision sur le découpage en scènes. Les descripteurs utilisés sont les informations du mouvement extraites directement du flux compressé, à savoir les vecteurs de mouvement et le nombre de macro-blocs codés en intra. À partir de ces informations, les auteurs déterminent trois autres mesures telles que :

- Le manque de mouvement,
- Les mouvements de caméra : panoramiques et zooms seulement et
- La présence des transitions vidéo de type cut.

Le manque de mouvement est caractérisé par un seuillage sur la valeur moyenne, μ, des modules des vecteurs de mouvement pour chaque image de type P. Cette moyenne est définie par :

$$\mu = \frac{1}{MN - I} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \sqrt{v_x^2(i,j) + v_y^2(i,j)} \tag{92}$$

$$\mu < S_{no-motion} \tag{93}$$

avec M et N les dimensions d'une image, I le nombre de macro-blocs codés en intra, v_x et v_y sont les coordonnées du vecteur de mouvement et $S_{no-motion}$ un seuil fixé à 4 par les auteurs.

L'estimation du mouvement de la caméra est obtenu par l'estimation des paramètres de mouvement relatif au zoom et au panoramique horizontal. L'estimation est obtenue à l'aide de la méthode des moindres carrés appliquée aux images de type P. Enfin les auteurs proposent un détecteur des transitions vidéo basé sur l'analyse des variations du nombre de macro-blocs codés en intra. Pour ce faire, les auteurs calculent la différence des valeur de μ entre deux images de type P consécutives. Cette différence est donnée par :

$$\Delta\mu(k) = \mu(k) - \mu(k-1) \tag{94}$$

avec $\mu(k)$ la valeur de μ définit par l'équation 92 pour l'image de type P et d'index k.
Ainsi dans le cas d'un transition vidéo abrupte les valeurs de $\Delta\mu(k)$ augmentent fortement. Pour améliorer la détection, les auteurs combinent la mesure précédente avec la mesure suivante :

$$Cut(k) = Intra(k) + \beta\Delta\mu(k) \tag{95}$$

avec $Intra(k)$ le nombre de macro-blocs codés en intra dans l'image de type P et d'indice k et β un coefficient de pondération.

Enfin la décision finale est prise par rapport à l'équation (95). Si les valeurs de $Cut(k)$ sont supérieures à un seuil S_{cut} alors l'image d'indice k est étiquetée comme un cut. Pour les travaux présentés dans ce papier les auteurs ont fixés β à 10 et S_{cut} à 400. À présent, les auteurs proposent de caractériser les scènes qui comportent des évènements importantq dans les matchs de football. Ces évènements sont les coups de pied de coin (corner), coups francs et bien entendu les buts. Pour cela les auteurs ont modélisé ces événements par des automates d'état finis donnés par les figures 29 et 30. Les résultats des ces détections d'événements sont

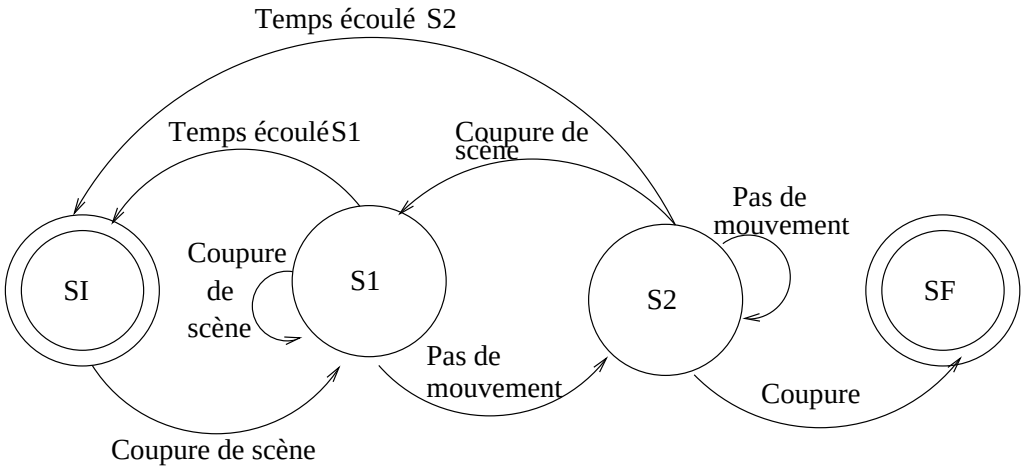


FIG. 29 – Automate de détection des corners et coups francs

très satisfaisants dans le cas de la détection des buts mais nettement moins bons en ce qui concerne la détection des corners et des coups francs. Malheureusement, les auteurs ne nous renseignent pas sur la durée du corpus vidéo utilisé ainsi que sur les performances en termes de rappel et précision.

Dans la suite de ce papier les auteurs présentent une autre approche utilisant les modèles de Markov cachés pour l'indexation des contenus multimédia. La méthode utilisée traite chacun des flux audio et vidéo indépendamment. Le flux audio est découpée en fenêtre d'analyse puis

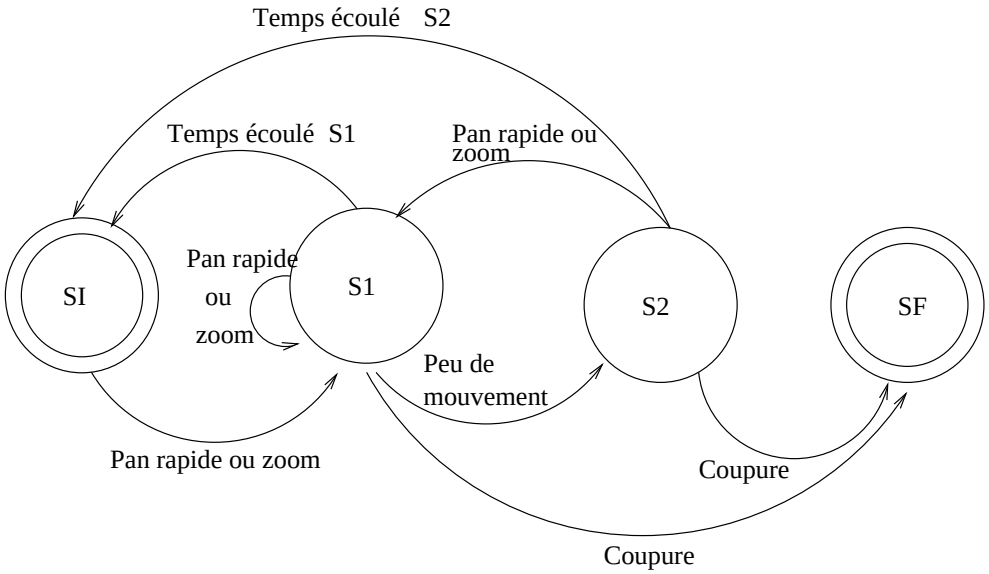


FIG. 30 – Automate de détection des buts

pour chaque fenêtre un vecteur de descripteurs tels que les Mel-Cepstrum coefficients (MFCC) ou bien le taux de passage par zéro (ZCR) est calculé. De même pour le flux vidéo, où pour chaque couple d'images adjacentes un vecteur de descripteurs tels que les histogrammes de luminance, les vecteurs de mouvement et la différence pixel-à-pixel est calculé. Après quoi, les vecteurs de descripteurs des deux flux sont soumis à un modèle de Markov caché (HMM) qui va déterminer les frontières des scènes et classifier ces dernières selon qu'elles soient :

- de *dialogue* : le signal audio est principalement composé de parole et les changements de contenu vidéo sont alternés du type ABAB...
- d'*histoire* : le signal audio est principalement composé de parole et le contenu du flux vidéo est composé d'un leitmotiv de la forme ABCADEFAG...
- d'*actions* : Le contenu audio appartient principalement à une classe et n'est pas de la parole et le contenu vidéo est composé de plans dont le contenu est toujours différent du type ABCDEF...
- *générique* : Le contenu audio appartient principalement à une classe mais la contenu vidéo n'a aucune cohérence.

De plus, en utilisant les modèles de Markov cachés (MMC) les auteurs proposent aussi une détection des événements importants dans les matchs de football comme ceux présentés plus haut. Ainsi, les auteurs font appel à deux topologies différentes pour les MMC : bottom-up et top-down. Avec la topologie top-down, les auteurs proposent la détection des événements tels que les corners, les coups francs et les buts. Quant à l'autre topologie proposée, bottom-up, elle permet une segmentation en scènes à partir des descripteurs audio et vidéo extraits auparavant puis la classification de ces mêmes scènes selon les classes définies plus haut.

Les auteurs présentent largement les résultats obtenus par les méthodes à base de MMC. Les expérimentations ont été menées sur 2 heures de vidéo au format compressé MPEG-2 d'un match de football. En ce qui concerne la détection des événements marquants, les performances varient de 14 à 74% pour le rappel. Sachant que la détection des buts est bien plus robuste que

celle des coups francs. Les auteurs expliquent les faibles performances dans la détection des corners et des coups francs par le fait qu'il est très difficile de modéliser un scénario pour ces actions, contrairement aux buts. En ce qui concerne la classification des scènes, la topologie bottom-up donne de bons résultats. Les expérimentations ont été menées sur des documents multimédia de contenus variés : journal télévisé, programme de variété, documentaire, etc... Ces résultats sont donnés par la table 31. Bien que les auteurs ne nous renseignent pas sur

Type de scène	Rappel	Précision
Dialogue	81%	64%
Action	79%	74%
Histoire	76%	69%
Générique	69%	67%

FIG. 31 – Résultats classification des scènes [LM00]

la durée effective de leur base de données vidéo de test, nous pouvons dire que les résultats présentés dans ces travaux sont tout de même très prometteurs.

Boccignone, De Santo et Percannella [BSP99]

Dans [BSP99], les travaux présentés traitent d'une approche cross-média pour l'indexation des vidéos dans le domaine compressé. La méthode consiste à traiter les flux audio et vidéo compressés de manière parallèle puis à fusionner le résultat des analyses précédentes afin de caractériser les frontières de scène.

Tout d'abord, l'analyse vidéo consiste à extraire un vecteur de descripteurs pour chaque image. La taille de ce vecteur est égale au nombre de macro-blocs contenus dans l'image et le nombre de bits nécessaires au codage de chacun de ces macro-blocs est utilisé comme descripteur. Ainsi, le vecteur de descripteurs vidéo est de la forme :

$$F_t = (f_{11} \dots f_{ij} \dots f_{mn})_t^T \tag{96}$$

avec f_{ij} le nombre de bits nécessaires au codage du macro-bloc de coordonnées (i, j) dans l'image à l'instant t et m, n étant les dimensions en termes de blocs de l'image.
Après quoi, les auteurs définissent une métrique entre les images par l'intermédiaire du calcul d'une distance entre les vecteurs de descripteurs. Pour deux images successives, t et $t^{'}$, la mesure de distance s'exprime sous la forme :

$$d(F_t, F_{t'}) = \frac{\sum_{i,j} |f_{ij} - f_{ij}^{'}|}{B_t} \tag{97}$$

avec B_t le nombre total de bits nécessaires pour coder la totalité des macro-blocs de l'image à l'instant t.
Dès lors que $d(F_t, F_{t'})$ est supérieure à un seuil Th, l'image courante est considérée comme une frontière de plan de montage. Toutefois, pour accroire la robustesse de cet algorithme les auteurs préconisent d'utiliser un vecteur de descripteurs compensé, $\widehat{F}_t$. Les composantes de ce nouveau vecteur compensé sont maintenant de la forme :

$$\widehat{f}_{ij} = f_{ij} + w_{DC} f_{ij}^{DC} \tag{98}$$

avec f_{ij}^{DC} la somme des six coefficients DC du macro-bloc de coordonnées (i, j) et w_{DC} un facteur pondérateur fixé de manière expérimentale à 0.8 par les auteurs.

Concernant l'analyse du flux audio, les auteurs proposent de modéliser le signal de la manière suivante :

$$a(t) = w_s s(t) + w_m m(t) + w_n n(t) + w_\sigma \sigma(t) \tag{99}$$

Cette modélisation peut être interprétée comme étant une combinaison linéaire pondérée des différentes classes de son telles que la parole (s), la musique (m) et le silence (σ) et w_k des coefficients associés à chacune des composantes m, s ou σ.
Le codage audio MPEG découpe le flux audio en sous-bandes, puis applique une banque de filtres afin d'estimer pour un temps donné t les données contenues dans la sous bande d'index i, que l'on note $S[i]$. Tout cela peut-être formellement écrit de la manière suivante :

$$S_t[i] = \sum_{n=0}^{N} a[t-n] h_i[n] \tag{100}$$

avec N le nombre d'échantillons par fenêtre d'analyse et h_i le filtre d'index i.
La méthode d'analyse du flux audio est décomposée en deux étapes. La première étape consiste à découper le flux audio en fenêtres de taille fixe puis à extraire un ensemble de descripteurs pour chacune de ces fenêtres. Chacune des fenêtres est alors classifiée en parole, musique, silence au cours de la seconde étape.
Les descripteurs calculés sont la valeur moyenne des échantillons (voir équation 101) de chaque fenêtre d'analyse ainsi que les quatre successifs premiers moments statistiques centrés (voir équation 102).

$$M_1 = \frac{1}{N} \sum_{k=1}^{N} x_k \tag{101}$$

$$M_{Ci} = \frac{1}{N} \sum_{k=1}^{N} (x_k - M_1)^i \ ,i = 2, \ldots, 5 \tag{102}$$

avec N le nombre total d'échantillons dans chacune des fenêtres, M_1 la valeur moyenne des échantillons, M_{Ci} le moment statistique centré d'ordre i et x_k l'échantillon d'indice k.
Dans le cadre de ces travaux, les auteurs ne considèrent que 8 sous-bandes sur les 32 que compte le format MPEG-1 Layer 2. Ainsi, chaque fenêtre est caractérisée par un vecteur de 40 descripteurs : 5 descripteurs pour chacune des 8 bandes considérées. Ce vecteur peut être écrit de la manière suivante :

$$F_t^* = (f_{11}^*, \ldots, f_{15}^*, \ldots, f_{81}^*, \ldots, f_{85}^*)_t^T \tag{103}$$

Concernant la seconde phase, la classification est réalisée à l'aide d'un réseau de neurones artificiel. Malheureusement, les auteurs ne donnent que très peu d'informations sur ce procédé : méthode d'entraînement, etc ...
La figure suivante schématise le processus complet d'analyse et d'indexation du flux audio : Un des points important de l'indexation cross-média est bien évidemment la fusion de l'information. Les auteurs proposent donc une méthode consistant à localiser de manière chronologique les frontières de classes audio, dès lors la méthode recherche la transition vidéo la plus proche

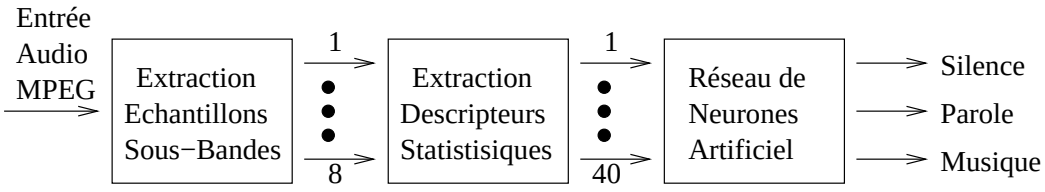


FIG. 32 – Schéma du processus de traitement du flux audio

en terme de distance. Si cette distance minimale est inférieure à un seuil donné, alors l'image de la vidéo courante est considérée comme une frontière de scène.

Les tests ont été menés sur une base de données de 15 vidéos au format MPEG composées de films, de journaux télévisés et de publicités. Pour la classification audio, les auteurs ont utilisé un réseau composé de 40 neurones d'entrée correspondant aux 40 descripteurs audio, de 2 couches cachées et de 3 neurones de sorties correspondant aux 3 classes audio considérées. La classe bruit a été considérée comme la classe de rejet, ce qui signifie que si une fenêtre n'appartient ni à la classe silence, ni à la classe musique elle est automatiquement classifiée en tant que bruit. Les performances annoncées sont très bonnes, le rappel est de l'ordre de 96% pour la détection des scènes sur l'ensemble du corpus. On peut déplorer l'absence du taux de fausses alarmes relatif à la détection des frontières de scènes. Enfin, les auteurs proposent de supprimer l'utilisation des différents seuils au cours de la méthode au profit de méthodes de décision statistique.

Chaisorn, Chua, Koh, Zhao, Xu, Feng et Tian[CCK+03]

Dans [CCK+03], la méthode proposée a été utilisée pour participer à la campagne d'évaluation TRECVID 2003 concernant la segmentation en scènes des journaux télévisés. Dans ces travaux les auteurs proposent d'utiliser une combinaison de descripteurs audio et vidéo afin de caractériser les frontières de scènes. Puis, les auteurs proposent de classifier les scènes précédemment détectées en deux classes : scènes de reportage, autres types de scènes. La figure 33 illustre le fonctionnement global de la méthode proposée ici : La première étape consiste à détecter les frontières de plans de montage, pour cela les auteurs utilisent une méthode d'analyse multi-résolution développée dans [LKC00]. Cette méthode affiche de bonnes performances tant dans la détection des transitions abruptes que graduelles. Une fois les frontières de plans de montage caractérisées, les auteurs proposent une classification de ces plans de montage suivant 12 catégories : Introduction, Présentateurs, Deux présentateurs, Foule, Interview/Parole, Reportage en direct, Sport, Contenant du texte, Spécial, Finance, Météo et Publicité. Malheureusement, les auteurs ne donnent pas plus de précisions sur les définitions des ces catégories. Les auteurs présentent, ensuite, un ensemble de descripteurs audio et vidéo nécessaire à la classification des plans de montage et des scènes. Ces descripteurs sont regroupés en trois grandes familles :

1. Descripteurs bas-niveau
   - Histogrammes de couleur : Modélisent la composition visuelle d'un plan. Utilisés pour la détection des plans tels que les plans de météo, de finance, etc ...
2. Descripteurs temporels

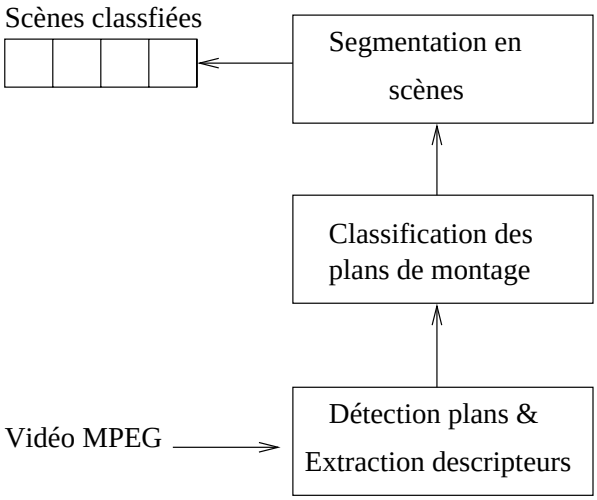


FIG. 33 – Schéma de la méthode complète

- Changement de scène : Indique le numéro des images correspondant aux frontières de scènes. Obtenu par la calcul de la différence des histogrammes associés aux images clés de deux plans consécutifs.
- Audio : Descripteur important pour la détection des plans de sport et les plans d'introduction. Dans le cas de sport, le flux audio est composé à la fois de parole (commentaires) et de bruit de fond. Dans le cas de plans d'introduction (génériques), la parole est associée à de la musique de fond.
- Activité de mouvement : Le mouvement est classifié en quatre niveaux tels que nul, bas, moyen et haut.
- Durée d'un plan de montage : Ce descripteur est à la fois utilisé pour la classification des plans de montage mais aussi des scènes.

3. Descripteurs haut-niveau basés objet
   - Visage : Pour chaque plan de montage, les auteurs extraient le nombre de visages présents ainsi que leur taille. Ce descripteur est utilisé dans la classification des plans, plus particulièrement pour caractériser les plans qui contiennent un ou deux présentateurs.
   - Type de plan : Les auteurs proposent trois types de plans : les plans rapprochés, les plans moyens ou les plans éloignés.
   - Texte à l'écran : Une scène-texte est une scène qui contient principalement du texte centré au milieu. On rencontre généralement ce type de scènes à la fin des rencontres sportives pour afficher le résultat et le résumé de la rencontre.
   - Phrase d'amorce : Les auteurs extraient la phrase d'amorce caractérisant généralement le début d'une scène de journaux télévisés. Les auteurs recherchent aussi si, pour chaque plan de montage détecté, cette phrase est présente ou non.

La phase suivante (voir schéma 33) consiste à classifier les plans de montages détectés. La règle de décision pour laquelle les auteurs ont opté est basée sur un arbre de décision. Le vecteur de descripteurs utilisé est de la forme :

$$S_i = (a, m, d, f, s, t, c)^T \tag{104}$$

avec a la classe audio, m le niveau d'activité de mouvement, d la durée du plan, f le nombre de visages détectés, s le type de plan, t le nombre de ligne de texte et c un booléen relatif à la présence ou non de texte à l'écran.

Concernant la détection des frontières de scènes ainsi que la classification de ces mêmes scènes, les auteurs proposent une méthode basée sur l'utilisation des modèle de Markov cachés. Un plan est modélisé par son type, un booléen relatif au changement ou non de lieu et la présence ou non d'une phrase d'amorce. Cette modélisation peut être formulée de la manière suivante :

$$S = [t, l, c] \tag{105}$$

avec t le type de plan, l le booléen de changement ou non de lieu et c le booléen de présence ou non de la phrase d'amorce.
À partir des composantes du vecteur S (voir équation 105), il est possible de former 64 différents vecteurs car t peut prendre 17 valeurs différentes, l 2 et c 2 d'où $17 \times 2 = 34$.

Pour la classification des scènes, les auteurs utilisent des heuristiques. Ils prennent en considération le type de plan de montage détecté, ainsi que des règles relatives à chacune des deux classes de scènes considérées. Ainsi, ils classifient les scènes détectées préalablement en scènes de journaux télévisés ou en autre type de scène.

Les tests ont été menés sur le corpus de test de TRECVID 2003 comprenant 120 heures de journaux télévisés diffusés sur CNN et ABC durant l'année 1998. Les auteurs ont utilisé les 60 premières heures de la base de données pour entraîner le système et le reste pour les tests. Les résultats sont, concernant la segmentation en scènes, de l'ordre de 70% pour le rappel et 74% pour la précision. En ce qui concerne la classification des scènes en scènes de journaux télévisés, le rappel est égal à 93% et la précision à 92%. Ces résultats attestent de la robustesse de la méthode utilisée car ces performances ont été obtenues sur un corpus de grande taille. Toutefois, les travaux présentés ici sont focalisés uniquement sur les journaux télévisés.

Tsekeridou et Pitas [TP99]

Dans les travaux de [TP99], les auteurs proposent une méthode d'indexation des contenus multimédia par segmentation des flux audio et vidéo. Ces travaux concernent le traitement des flux audio génériques, sans a priori sur le contenu contrairement à [NAT98] qui s'intéressent aux scènes de violence ou bien à [NK97] qui se préoccupent de l'indexation des journaux télévisés. L'objectif de cette indexation consiste à identifier les plages de silence dans le flux audio.

Le traitement audio commence par l'extraction de descripteurs bas-niveau tels que les coefficients MFC utilisés pour la segmentation parole/musique. La détection des silences est assurée par l'analyse de l'énergie ainsi que du taux de passage par zéro qui permet par la même occasion de caractériser les zones de paroles. Le calcul des descripteurs audio bas-niveau est obtenu avec des fenêtres d'analyse d'une durée de 10ms. Pour différencier les zones de silence des zones de parole, les auteurs définissent, d'après les travaux de [RS78], des seuils en relation avec les fonctions de moyenne pour l'énergie, $M_{t,n}$, et pour le taux de passage par zéro, $Z_{t,n}$. Concernant l'énergie, deux seuils ont été définis et donnés par les équations suivantes : un

seuil haut (équation 106) et un seuil bas (équation 107).

$$M_{thr,up} = \frac{E[M_t] - M_{thr,low}}{2} \quad (106)$$

$$M_{thr,low} = max(M_{t,n}) \quad (107)$$

avec $E[M_t]$ l'espérance mathématique de M_t, la valeur de l'énergie pour la fenêtre d'analyse d'indice t.

Après avoir terminé avec la segmentation en Parole/Silence, la tâche maintenant consiste à caractériser les zones voisées des zones non voisées dans les segments de paroles précédemment détectés. Le principe de la méthode utilisée pour cette segmentation est basé sur le fait que les parties non voisées contiennent de manière significative des hautes fréquences contrairement aux parties voisées. Contrairement à ce que nous avons vu dans la section 3.1.3 et démontré par les figures 9 et 8, le taux de passage par zéro est un très bon candidat pour la caractérisation des parties voisées et non voisées. Or, dans ces travaux les auteurs, s'inspirant des travaux présentés dans [RS78], utilisent plutôt la proportion d'énergie dans les bandes de hautes et basses fréquences du spectre. Les bandes ont une fixe taille de 2kHz et la fréquence d'échantillonnage est de 22kHz. La décision est prise par rapport à un seuil sur le rapport entre la proportion de l'énergie dans les hautes fréquences et la proportion dans les basses fréquences. Si ce rapport est supérieur à 0.25 alors la zone considérée est étiquetée comme une zone non-voisée et ignorée pour la suite du traitement au même titre que les zones de silence localisées auparavant.

Les tests ont été menés sur des vidéos dont la bande sonore est échantillonnée à 22kHz et les échantillons codés sur 16 bits. Les auteurs annoncent de bonnes performances pour la segmentation en Parole/Silence. Toutefois, quelques zones ont été classifiées comme étant des zones non-voisées au lieu de silence. En revanche la réciproque n'est pas vraie, aucune zone de silence n'a été étiquetée comme non voisée. Malheureusement, les auteurs ne fournissent pas de performances chiffrées en terme de rappel et de précision. D'autre part, le corpus de test utilisé pour illustrer cet article se compose de deux vidéos d'une durée de 6 et 4 minutes ce qui ne permet pas de valider une telle méthode. Cependant, les auteurs affirment avoir fait d'autres expérimentations avec un corpus plus conséquent.

Conclusion

Ainsi dans ce chapitre, nous avons étudié diverses méthodes d'indexation des documents vidéo par l'analyse cross-média. Cette étude bibliographique a montré à quel point il est évident que les approches cross-médias sont les plus performantes car elles font intervenir l'analyse des deux flux : audio et vidéo. Ces méthodes exploitent ainsi pleinement l'information contenue dans un document audiovisuel.

Deuxième partie

Fusion de l'information cross-média

Chapitre 5

Modèle des scènes audiovisuelles dans les contenus télédiffusés

Ce chapitre présente successivement la problématique, le cadre général utilisé pour l'analyse et l'indexation des flux télédiffusés. Enfin, nous présentons le modèle que nous proposons.

5.1 Problématique

Dans le chapitre 2, nous avons pu voir l'intérêt du problème d'indexation des documents audiovisuels en scènes. Il est d'autant plus flagrant qu'à ce jour il existe le standard de description MPEG-7 [Mpe, SS01]. MPEG-7 propose un cadre normalisé pour la description des contenus, assurant ainsi l'interopérablité des systèmes multimédia : l'utilisateur peut aussi bien accéder au contenu via Internet, DVR (Digital Video Recorder) ou encore via son téléphone portable. Néanmoins, comme tous les standards, MPEG-7 ne normalise pas la *façon de faire*, d'où la diversité des approches citées dans les chapitres 2 et 3.

Dans cette partie, nous proposons à l'attention du lecteur la première contribution de cette thèse. À savoir, le modèle et la méthode de fusion des résultats des détecteurs audio et vidéo dans le problème du *chapitrage sémantique* des contenus audiovisuels télédiffusés. L'ensemble des travaux présentés s'inscrivent dans le cadre du projet CASSANDRA [Cas] mené au Philips Research Natlab (NL). Dans ce projet, Philips développe des algorithmes d'analyse et d'indexation des contenus multimédia.

Dans le chapitre 5, nous introduisons le cadre général d'analyse audio-vidéo et nous présentons le modèle de scène considéré. La chapitre 6 contient le modèle de fusion des résultats d'analyse des flux.

Dans le cadre de cette thèse, nous n'avons pas travaillé sur l'analyse vidéo, le détecteur des changements des plans du projet CASSANDRA [Cas] ayant déjà été développé. Bien qu'ayant été utilisateur des détecteurs audio du projet CASSANDRA, nous avons élaboré nos propres outils d'analyse du flux audio qui sont présentés dans le chapitre 7.

5.2 Cadre général d'analyse des flux télédiffusés

Dans cette thèse, nous nous intéressons à l'analyse temps réel des flux télédiffusés très hétérogènes qui proviennent des chaînes de distribution (satellite, cable, hertzien). L'objectif fixé est ambitieux puisqu'il consiste à proposer un modèle générique de scène audiovisuelle pour l'ensemble de ces contenus. Compte tenu de la variété des contenus un tel modèle ne peut pas donner entière satisfaction, néanmoins tel était l'objectif fixé dans le cadre du projet CASSANDRA[Cas] et par conséquent de cette première partie de thèse.

Dans cette partie, nous présentons le cadre général, développé au sein du projet CASSANDRA, de l'analyse des flux audiovisuels. Le schéma de la figure 34 illustre le fonctionnement général du processus d'analyse et d'indexation. Ce processus d'analyse consiste à caractériser les plans de montage vidéo, dans le flux vidéo, et les frontières des silences, dans le flux audio. Tout cela est effectué à partir des descripteurs bas niveau extraits des flux audio et vidéo dans les domaines compressés et non compressés.

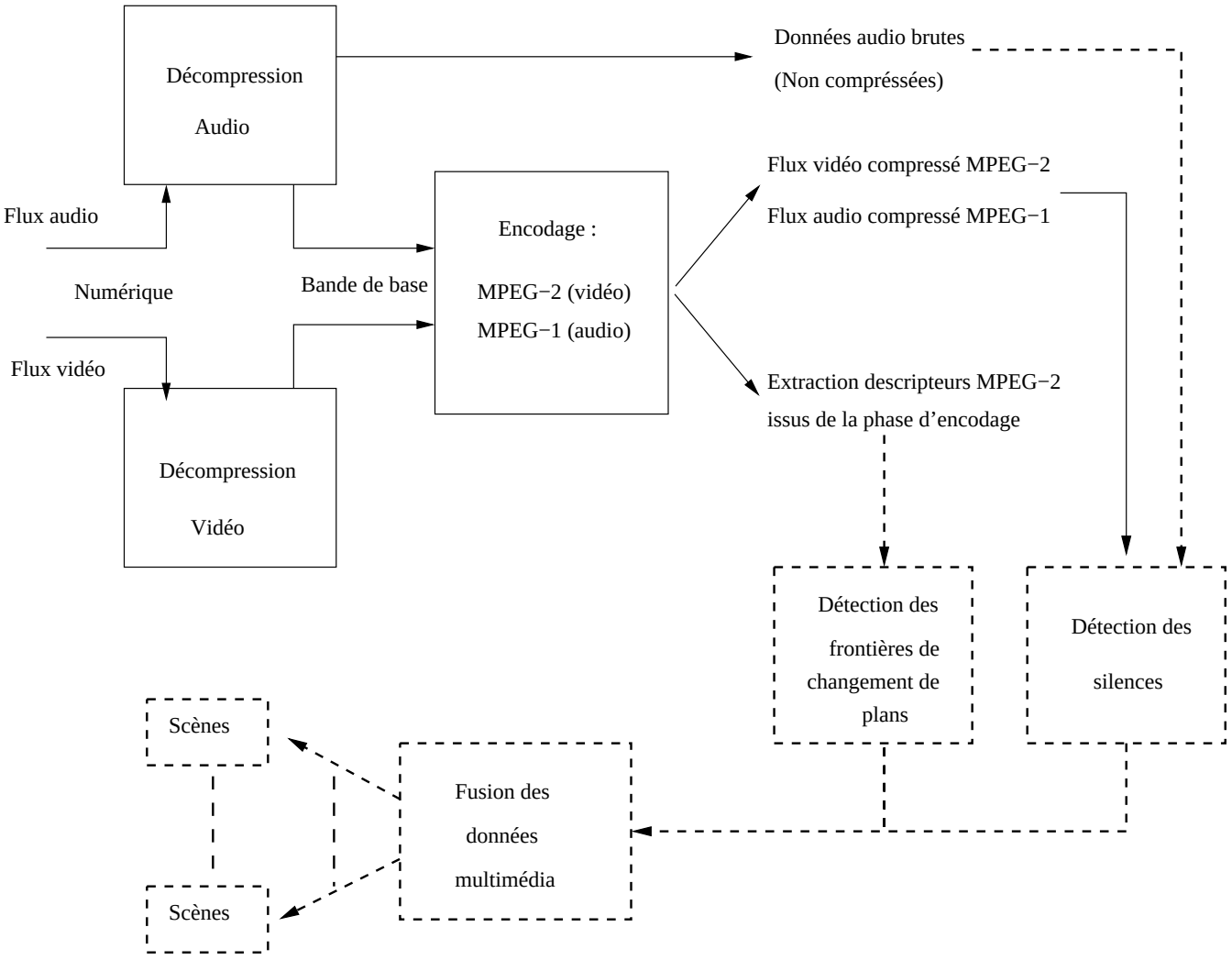


FIG. 34 – Diagramme du processus d'analyse et d'indexation des flux audiovisuels

Dans ce schéma, nous avons indiqué les blocs dans lesquels nous avons apporté notre contribution (traits en pointillés gras).

Comme il possible de le remarquer sur le schéma de la figure 34, le projet CASSANDRA[Cas] exploite les descripteurs audio et vidéo bas niveau extraits du flux brut (audio) ou lors de l'encodage du flux compressé (audio et vidéo). Ces descripteurs sont utilisés dans l'objectif de l'indexation du flux en scènes. Ainsi, nous allons maintenant définir la notion et le modèle d'une scène d'un contenu télédiffusé tels qu'ils ont été considérés dans cette étude.

5.3 Modèle de scène pour les contenus audiovisuels télédiffusés

Cette section est consacrée à la présentation du modèle de scène que nous proposons de caractériser dans les contenus audiovisuels, par la méthode d'indexation cross-média ainsi qu'à la définition d'une mesure temporelle entre les transitions audio et vidéo.

5.3.1 Définition

Tout d'abord, dans notre définition des scènes nous tenons compte d'une large variété de contenus multimédia télédiffusés. Nous ne cherchons pas à proposer une solution pour des contenus avec des contraintes a priori fortes telles que les programmes sportifs, journaux télévisés, etc... Au contraire, nous nous intéressons au large spectre du contenu que nous qualifions d'*artistique* : films, séries télévisées, documentaires, etc...

Comme le montre la figure 1, les contenus audiovisuels que nous appelons *artistiques*, par opposition aux contenus sportifs, admettent une structuration hiérarchique.
De nombreux travaux proposent des modèles de scène comme un regroupement de plans de montage successifs en fonction du contenu [RS03, MHI+02, KCL00] (cf chapitre 2).
De surcroît, d'autres travaux considèrent un modèle spécifique de scènes, comme les scènes de dialogues [AAW01, DCBP05] ou bien les scènes d'intérieur et d'extérieur [MHI+02].
A l'opposé dans nos travaux [NLBP+04], nous cherchons à proposer un modèle de scène générique sans a priori sur le type de scène et sur le genre du contenu.
Le contenu d'une scène véhicule un seul et même message *sémantique*.

Le problème que pose la notion de scène est de trouver une bonne définition du terme *sémantique* car cette notion est très subjective. C'est pourquoi nous avons sélectionné un ensemble de sujets ayant une activité liée au domaine de l'indexation vidéo donc ayant déjà travaillé sur des contenus mutlimédia variés provenant des mêmes sources de télédiffusion que celles utilisées pour la constistution de notre corpus. Nous avons interrogé cet ensemble d'individus sur la définition qu'ils proposeraient d'une scène en fonction du type de contenu. Les personnes sont des hommes et des femmes dont la moyenne d'âge est d'environ 25 ans, ils représentent la tranche de la population qui regardent le plus vaste panel de contenus télédiffusés. Enfin, compte tenu du coût d'une telle enquête, le nombre de sujets interrogés a été limité à 4. Les tableaux des figures 35 et 36 résument les avis recueillis par cette population. Les termes tels que "individu", "la cohérence du lieu", "événements", etc... sous-entendent la similarité ou la conservation du terme considéré.

Nous avons collecté ces informations dans le but d'induire une définition globale de la notion de scène dans les contenus audiovisuels. D'après les informations recueillies, nous pouvons

Genre Vidéo \ Nom des sujets	Sujet 1	Sujet 2
Films	Lieu et Événement	Lieu
Journaux	Sujets et Lieu	Locuteurs et Reportages
Sports	Événements et Position caméra	Événements
Séries	Personnages	Personnages et Lieu
Dessins animés	Lieu et Personnages	Lieu
Documentaires	Action du sujet et Lieu	Reportage et Action du sujet
Variétés	Invités et Événements	Invités et Événements

FIG. 35 – Interprétation de la notion de scène par des sujets hommes selon différents types de contenus audio-vidéo

Genre Vidéo \ Nom des sujets	Sujet 3	Sujet 4
Films	Lieu	Lieu
Journaux	Locuteurs et Reportages	Locuteurs et Reportages
Sports	Pas d'idée	Événements
Séries	Lieu	Lieu
Dessins animés	Lieu	Lieu
Documentaires	Pas d'idée	Reportage
Variétés	Pas d'idée	Événements

FIG. 36 – Interprétation de la notion de scène par des sujets femmes selon différents types de contenus audio-vidéo

formuler les quelques définitions suivantes pour les scènes en fonction des différents genres de contenus vidéo :

- **Définition 1** : La cohérence du lieu peut définir une scène dans les vidéos de type **films, séries et dessins animés**. Les **journaux** peuvent aussi être considérés mais avec un niveau de priorité inférieur. La cohérence du lieu signifie la localisation physique, en d'autres termes le lieu filmé par la caméra et non pas le lieu où se situe la caméra.
- **Définition 2** : L'alternance entre les reportages et les apparitions du présentateur doit permettre de caractériser les scènes dans le cas des **journaux télévisés**.
- **Définition 3** : Les événements importants tels que les buts, les coups de pied de surface, les pauses, etc... présents dans les manifestations sportives (en particulier dans les cas des matchs de football) sont de bons candidats pour matérialiser les frontières de scènes dans les **programmes sportifs**.
- **Définition 4** : Les différentes actions d'un personnage (apparition et/ou disparition) peuvent aussi être un bon critère dans le cas des vidéos de type **séries documentaires** et **variétés**. Dans le cas d'une émission de **variétés (jeu)**, lorsqu'un participant perd, il doit quitter le plateau et éventuellement un nouveau participant apparaît. Cette action peut signifier la fin d'une scène et le début d'une autre.
- **Définition 5** : Pour un type particulier d'émissions de **variétés** comme les **jeux télévisés**, le déroulement du jeu en lui-même est souvent composé de plusieurs manches. Dans ce cas précis, chaque manche peut être assimilée à une scène.

Comme le montrent les tables des figures 35 et 36, il est difficile de caractériser une scène

sans des méthodes de reconnaissance de la parole, de la musique, de détection d'objets spécifiques, d'objets d'intérêts. À titre d'exemple, si nous considérons le cas du changement de locuteurs, cela signifie qu'il faut à la fois détecter les changements de locuteurs dans la bande audio et dans le flux vidéo. Et, dans le domaine de la vidéo il faut faire appel à la détection et à la reconnaissance de visages. Ces réalisations font appel à des méthodes et modèles assez complexes qui, de surcroît, nécessitent de lourdes ressources de calcul. De plus, des règles spécifiques adaptées à chaque type de genre vidéo sont nécessaires.

C'est pourquoi, nous avons décidé de proposer un modèle qui est un bon compromis sur le plan de la simplicité, de la robustesse et de son aspect générique, i.e sans a priori sur le contenu. Ainsi, nous ne nous focaliserons pas sur la scène elle-même mais sur ses frontières. Nous considérerons qu'une scène est définie par deux frontières successives. Par conséquent, nous associons au modèle proposé une série d'heuristiques telles que :

- En vidéo, les différents types de transitions entre les plans telles que les fondus ou les effets de balayages sont utilisés pour véhiculer un message au spectateur. Certains peuvent supposer que ces types de transitions progressives permettent de lier de manière sémantique les plans entre eux. Au contraire, les transitions vidéo abruptes - *cuts* - signifient une coupure *sémantique* également. Nous allons donc supposer *que seuls les cuts sont utilisés pour caractériser une frontière de scène vidéo.*
- Dans le domaine audio, un changement de scène intervenant en plein milieu d'une conversation ou bien encore d'un morceau de musique n'est pas réalisable. C'est pourquoi, nous pensons qu'il est raisonnable de considérer *la majorité des frontières de scène audio comme étant matérialisées par des silences.*

En conclusion de cette étude, nous avons décidé de combiner les transitions abruptes de types coupures (*cuts*) dans le flux vidéo avec la présence de silence dans le flux audio pour localiser les frontières de scènes dans les documents audiovisuels.
Le modèle de frontière entre deux scènes représente la coïncidence d'un silence dans le flux audio avec une transition cut dans la vidéo.
Il s'agit de localiser les zones de correspondance temporelle entre les coupures dans le flux vidéo et les silences dans le flux audio. De plus, comme nous travaillons sur des contenus artistiques, le multiplexage entre les flux audio et vidéo est issu d'une opération de post-traitement qui s'appelle le *montage*. Lors du montage, les auteurs désynchronisent volontairement et très légèrement les flux audio et vidéo afin d'adoucir les transitions. Ce qui signifie que le contenu audio de la scène suivante est audible quelques dixièmes de seconde avant (ou après) le contenu vidéo associé. De cet effet d'édition se traduit un léger décalage temporel dans les flux entre les coupures vidéo et les silences audio dans le cas d'un changement de scène. Pour pallier ce problème, nous avons défini une mesure signée de cette différence temporelle que nous avons appelée *jitter* et dont nous détaillerons les spécificités dans la section suivante.

5.3.2 Mesure de l'écart temporel : Jitter

Avant d'introduire la mesure de l'écart temporel entre les silences et les cuts nous devons définir une unité de temps commune pour les deux flux. Dans le flux audio, la classification des silences se fait par des fenêtres temporelles de quelques millisecondes : fenêtres de 512 échantillons audio avec une fréquence d'échantillonnage de 44100 Hz ce qui signifie que les fenêtres d'analyse ont une durée de 10 ms. Dans le flux vidéo, les transitions de type cuts

surviennent entre deux images consécutives. Nous allons donc considérer comme unité de temps l'écart temporel entre deux images : 1/25 ou 1/30 de seconde suivant le standard vidéo PAL/SECAM ou NTSC (voir figure 37a.
La graduation de l'axe temporel représente le numéro des images depuis le début du document audiovisuel. Jitter sera donc considéré comme une statistique qui nous permettra de fusionner les résultats audio et vidéo.

Considérons maintenant $v(i)$ le numéro de la première image après la dernière transition dans le i-ième plan de montage du flux vidéo entier qui en contient I (plans de montage), donc $i \in [1, \ldots, I]$. Considérons maintenant $(ab(s), ae(s))$ un couple de marqueurs temporels correspondant aux frontières de début et de fin du s-ième silence du flux audio qui en contient S, donc $s \in [1, \ldots, S]$. D'après la définition de $ab(s)$ et de $ae(s)$, nous pouvons énoncer la propriété suivante : $\forall s \in [1, \ldots, S]$, $ab(s) \leq ae(s)$. Pour un couple (i, s) donné notons $a = v(i) - ab(s)$ et $b = v(i) - ae(s)$. Ainsi, nous pouvons définir notre mesure J de la manière suivante :

$$J(i) = \begin{cases} 0 & \text{Si } \exists s^* \text{ tel que } ab(s^*) \leq v(i) \leq ae(s^*) \\ min_s(min(|a|, |b|)) \times sign(argmin(|a|, |b|)) & \text{Sinon} \end{cases} \tag{108}$$

avec $sign(x) = \begin{cases} 1 \text{ si } x \geq 0 \\ -1 \text{ sinon} \end{cases}$

L'illustration graphique de cette définition est présentée par la figure 37b.

Afin d'optimiser la rapidité des calculs du jitter pour la i-ième coupure dans le flux vidéo, il n'est pas nécessaire de considérer l'ensemble des S silences. Il suffit juste de considérer le silence le plus proche temporellement, qu'il soit avant ou après la transition vidéo $v(i)$ considérée. Cela correspond à la deuxième condition de la définition donnée par l'équation (108) au cas où la première ne serait pas respectée.

Nous nous sommes imposés une contrainte de traitement temps réel pour le calcul du jitter et la prise de décision concernant la détection des frontières de scène. Pour cela, nous allons utiliser une règle de décision statistique basée sur le modèle bayésien et sur le test du maximum de vraisemblance. Nous présentons les règles de décision dans le chapitre 6 suivant.

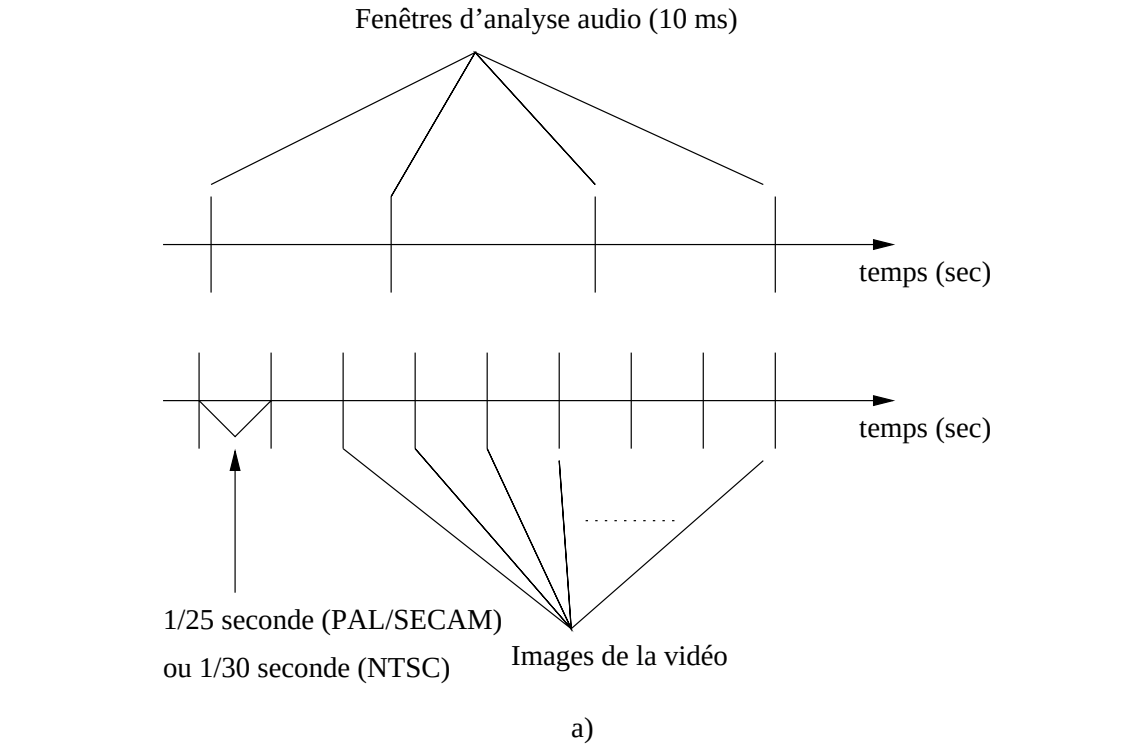


Fig. 37 – Schémas d'illustration de la définition du jitter : a) illustration de l'unité de temps et b) illustration du calcul du jitter

Chapitre 6

Modèle de fusion de l'information

Dans ce chapitre, nous présentons le modèle utilisé dans nos travaux pour la fusion de l'information [NLBP+04]. Travaillant sur des données réelles bruitées, il nous a semblé cohérent de proposer, pour notre règle de décision, le modèle classique de décision bayésienne au sens de maximum de vraisemblance des hypothèses émises. Ainsi, nous nous concentrons sur la formulation du schéma de décision appliqué à la variable aléatoire *jitter* définie dans le chapitre précédent.

Enfin, dans les annexes B.2 et B.3 nous proposons au lecteur les fondements mathématiques de cette approche.

6.1 Règles de décision sur les frontières de scènes

Dans le chapitre précédent, nous avons formulé notre modèle de changement de scène et sachant que des décalages temporels sont possibles nous avons introduit une mesure de décalage temporel que nous avons appelée *jitter*. Nous associons une variable aléatoire à cette mesure. Pour décider si notre mesure *jitter* observée pour un couple silence audio et coupure video dans le voisinage d'un instant de temps fixe correspond à une frontière de scène, nous proposons un schéma de décision bayesienne dans le cas à deux hypothèses. La première exprime le fait que la valeur observée de jitter correpond aux changements de scene et la deuxième exprime le contraire.

De manière générale, nous considérons X et H deux variables stochastiques associées aux évènements. $Pr(X)$ et $Pr(H)$ les probabilités relatives à ces événements et $Pr(X|H)$ et $Pr(H|X)$ leur probabilités conditionnelles. D'après le théorème de Bayes, nous avons :

$$Pr(H|X) = \frac{Pr(X|H) \times Pr(H)}{Pr(X)} \tag{109}$$

Si nous considérons maintenant deux réalisations, x et h, relatives à X et H l'équation (109) peut s'écrire sous la forme suivante [CH78] :

$$f(h|x) = \frac{f(x|h) \times f(h)}{f(x)} \tag{110}$$

avec $f(x)$ et $f(h)$ les fonctions de densité de probabilité de x et h, $f(x|h)$ et $f(h|x)$ les fonctions de densité de probabilités conditionnelles associées.

Dans notre problème, x correspond à toute nouvelle valeur de jitter mesurée à la volée correspondant à la variable X. Considérons deux hypothèses :

- H_1 : Jitter correspondant à une frontière de scène et
- H_2 : Jitter ne correspondant pas à une frontière de scène.

Nous admettons aussi que H_1 et H_2 réalisent une partition de l'ensemble des hypothèses. Donc,

$$\begin{aligned} Pr(H_1) + Pr(H_2) &= 1 \\ Pr(H_1 \times H_2) &= 0 \end{aligned} \tag{111}$$

Considérons les lois associées à ces hypothèses $f(h1)$ et $f(h2)$ respectivement.

Si l'on applique à notre problème la formule donnée par l'équation (110) nous obtenons :

$$\begin{cases} f(h_1|x) = \frac{f(x|h_1) \times f(h_1)}{f(x)} \\ f(h_2|x) = \frac{f(x|h_2) \times f(h_2)}{f(x)} \end{cases} \tag{112}$$

Selon le développement classique de la théorie de la décision bayésienne [CH78], notre but est de caractériser le maximum entre $f(h_1|x)$ et $f(h_2|x)$ dans l'équation (112).
Comme $f(x)$ ne dépend pas de h, il est équivalent de maximiser la fonction de vraisemblance :

$$L(h) = f(x|h) \times f(h) \tag{113}$$

ou de manière équivalente de prendre la valeur de $\max(L_1, L_2)$ avec :

$$\begin{cases} L_1 = f(x|h_1) \times f(h_1) \\ \text{et} \\ L_2 = f(x|h_2) \times f(h_2) \end{cases} \tag{114}$$

Si $L_1 > L_2$, alors l'hypothèse H_1 est validée, sinon c'est H_2 qui est validée.

Tout ceci est aussi équivalent à considérer le rapport de vraisemblance et de comparer ce dernier avec un seuil, $l \geq 1$ ($l = 1$ dans le cas trivial, en pratique il peut être ≥ 1). Nous pouvons formuler cette méthode de la manière suivante :

$$\begin{array}{l} \frac{L_1}{L_2} \; \begin{array}{l} > \\ < \end{array} \; l \; \begin{array}{l} \Longrightarrow H_1 \\ \Longrightarrow H_2 \end{array} \end{array} \tag{115}$$

La notation $\Longrightarrow H_i$ signifie que l'hypothèse H_i est validée.
Si l'on prend en compte la propriété donnée par l'équation 111, le rapport de vraisemblance peut être formulé de la manière suivante :

$$\frac{L_1}{L_2} = \frac{f(x|h_1) \times f(h_1)}{f(x|h_2) \times (1 - f(h_1))} \tag{116}$$

Dans la suite, nous supposerons que la probabilité, $f(h_1)$, d'un changement de scène est connue a priori. Cette valeur de probabilité peut être estimée lors de l'apprentissage du système

sur une base d'apprentissage. Pour ce faire, il faut comptabiliser manuellement, sur la base d'apprentissage, tous les cas de frontières de scènes. La valeur de $f(h_1)$ correspond alors au rapport entre le nombre de cas de changements de scènes et le nombre total de transitions vidéo abruptes observées. Désormais, nous notons P cette probabilité. De plus, l'équation (116) pourrait être simplifiée si nous considérions les événements relatifs aux hypothèses H_1 et H_2 équiprobables. Cependant, selon la nature du problème il nous semble convenable de ne pas considérer cette équiprobabilité des hypothèses.

Nous allons maintenant considérer que la variable x, associée au jitter, suit les distributions Normales, $N(\mu_1, \sigma_1^2)$ et $N(\mu_2, \sigma_2^2)$, associées aux hypothèses H_1 et H_2 respectivement. Ainsi l'équation (116) exprimant le rapport de vraisemblance s'écrira maintenant sous la forme suivante :

$$\frac{L_1}{L_2} = \frac{(2\pi\sigma_1^2)^{\frac{-1}{2}} \times exp\left(\frac{-(x-\mu_1)}{2\sigma_1}\right)^2}{(2\pi\sigma_2^2)^{\frac{-1}{2}} \times exp\left(\frac{-(x-\mu_2)}{2\sigma_2}\right)^2} \tag{117}$$

Suivant la démarche classique [CH78], nous allons maintenant considérer le logarithme du rapport de vraisemblance. Nous obtenons ainsi :

$$R = \frac{1}{2}\left[\left(log(\sigma_2^2) + \frac{(x-\mu_2)^2}{\sigma_2^2}\right) - \left(log(\sigma_1^2) + \frac{(x-\mu_1)^2}{\sigma_1^2}\right)\right] \begin{array}{l} > \\ < \end{array} log(l)+log((1-P)/P \begin{array}{l} \Longrightarrow H_1 \\ \Longrightarrow H_2 \end{array} \tag{118}$$

Nous nous plaçons dans le cas théorique, $l = 1$, d'où :

$$R = \frac{1}{2}\left[\left(log(\sigma_2^2) + \frac{(x-\mu_2)^2}{\sigma_2^2}\right) - \left(log(\sigma_1^2) + \frac{(x-\mu_1)^2}{\sigma_1^2}\right)\right] \begin{array}{l} > \\ < \end{array} log((1-P)/P) \begin{array}{l} \Longrightarrow H_1 \\ \Longrightarrow H_2 \end{array} \tag{119}$$

L'équation (119) correspond à la formalisation de notre règle finale de décision. Toutefois, il reste des paramètres inconnus, (μ_1, σ_1^2) et (μ_2, σ_2^2), à estimer. Il existe deux moyens possibles pour estimer ces paramètres. Le premier [Tch96] consiste à apprendre de manière statistique les paramètres à partir d'une base d'apprentissage étiquetée manuellement pour chacune des hypothèses H_1 et H_2. Nous supposons que les valeurs que peut prendre le jitter (voir équation (108)) suivent la loi Normale, donc les paramètres peuvent être estimés au sens de maximum de vraisemblance de la manière suivante (cf Annexe B.4) :

$$\widehat{\mu} = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{120}$$

et

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2 \tag{121}$$

avec N le nombre de valeurs du jitter récoltées dans les bases d'apprentissage.

Donc, une fois ces paramètres estimés pour chacune des classes considérées à partir de la base d'apprentissage et après avoir récupéré une nouvelle valeur de jitter il faut injecter toutes ces données dans l'équation (119) pour prendre la décision sur la présence ou non d'une frontière de scène. Dans cette première approche d'estimation des paramètres, ces derniers ne

sont pas mis à jour à chaque nouvelle valeur de jitter mesurée hors de la base d'apprentissage et les décisions sont prises *à la volée*.

La seconde possibilité envisagée pour l'estimation des paramètres consiste à mettre à jour ces données après chaque décision prise à partir d'une nouvelle valeur de jitter. De même, la valeur de la probabilité P est aussi mise à jour. Cependant, cette méthode qui est de plus en plus utilisée dans la littérature [WTH00] nécessite généralement des modèles de classes plus complexes. À savoir, des modèles à base de mélange de gaussiennes ou bien d'autres types de méthodes itératives qui ont l'inconvénient de n'être pas en temps réel. Néanmoins, cette piste est à envisager comme perspective des travaux présentés ici.

6.2 Fusion des données audiovisuelles

Ainsi, le test des hypothèses statistiques exposé dans la section précédente permet de traduire la *coïncidence* des évènnements audio et vidéo relatifs aux frontières des scènes et ainsi *fusionner* les données multimédia. Nous allons, en particulier, insister sur des contraintes temps réel associées à notre appliction.

La méthode que nous proposons se décompose en trois grandes étapes :

1. Calcul des descripteurs audio et vidéo,
2. Fusion des descripteurs audio et vidéo et
3. Post-traitement.

Ces trois grandes étapes sont présentées en détail dans les trois sous-parties suivantes respectivement.

6.2.1 Calcul des descripteurs audio et vidéos

À ce stade, le détecteur des frontières de scènes reçoit les informations issues de l'analyse des flux audio et vidéo (voir chapitre 5). Dès qu'une frontière de silence ou bien une frontière abrupte de plan de montage est caractérisée, elle est immédiatement stockée dans une mémoire tampon. L'ensemble des informations stockées est exprimé en fonction du numéro d'image de la vidéo. Ce processus se répète pour toute la durée du fichier (ou du flux) vidéo compressé MPEG-2.

6.2.2 Fusion des descripteurs audio et vidéo

Dans cette seconde étape, le détecteur de frontières de scènes accède à la mémoire tampon contenant le résultat de l'analyse des flux audio et vidéo. Comme précisé dans la sous-partie précédente, toutes les informations sont converties au même format. À savoir, exprimées en fonction du numéro d'image dans la vidéo.
Après quoi, le système peut commencer à calculer les différentes valeurs pour le jitter et prendre les décisions en conséquence. Afin de plaider en faveur de notre méthode statistique de décision, nous avons introduit un autre schéma de décision simple basé sur l'utilisation d'un seuil fixe. Les deux paragraphes suivants présentent ces deux schémas de décision.

Schéma basé sur l'utilisation d'un seuil fixe

Le détecteur de frontières de scènes recherche l'éventuelle présence d'une transition vidéo de type coupure, $sc(i)$, qui serait comprise dans le segment temporel $[ab(k)-TJ, ae(k)+TJ]$ avec $ab(k)$ et $ae(k)$ les frontières du k-ième silence et TJ une valeur de jitter fixe. Ce seuil correspond à une règle de décision simple. S'il y a une transition vidéo, $sc(i)$, qui coïncide avec une frontière de silence alors nous pouvons supposer que nous sommes en présence d'une frontière de scène audio-visuelle. L'algorithme marque alors la transition vidéo $sc(i)$ comme une frontière de scène audio-visuelle, notée $mms(l)$. Finalement, une scène audio-visuelle est caractérisée par deux frontières successives, ce qui est équivalent au segment temporel $[mms(l-1), mms(l)]$ avec $mms(0) = 0$.

Il est important de noter que cette approche a été proposée par J. Nesvabda dans son travail de thèse[1]. Dans notre travail, nous l'utilisons à titre de comparaison avec la méthode de décision statistique que nous avons proposée.

Schéma basé sur l'utilisation de la méthode statistique

La nouvelle valeur de jitter est mesurée en recherchant les frontières de silence les plus proches de la transition vidéo courante, le segmentbut étant de calculer la valeur minimale du jitter pour une transition vidéo donnée. Les paramètres statistiques des deux classes "changement de scène" et "non-changement de scène" ayant été appris lors de l'apprentissage, la décision est prise suivant le modèle présenté dans la section 6.1.

6.2.3 Post-traitement

Une fois que le processus de détection des frontières de scène est terminée pour les deux schémas proposés, nous considérons la longueur de la scène détectée afin d'optimiser les performances de détection en affinant les résultats obtenus. De manière expérimentale, après différentes observations sur notre base de données vidéo, nous pouvons dire qu'une scène ne peut pas avoir une durée inférieure à 2 secondes. Ainsi, avec une telle durée minimale nous pouvons éliminer les quelques scènes sur-détectées correspondant à un changement de programme ou bien lorsque deux frontières de scènes sont trop proches dans le temps : par exemple, dans le cas d'une image noire entre deux publicités. Donc, le détecteur élimine la frontière de scène $mms(l)$ si la condition suivante est vérifiée :

$$mms(l) - mms(l-1) \leq 2 \times NISV \tag{122}$$

avec $mms(l)$ le numéro d'image dans la vidéo correspondant à la frontière de scène d'indice l et $NISV$ le nombre d'images par seconde de la vidéo.

6.3 Résultats et expérimentations

Diverses expérimentations ont été réalisées afin de valider notre modèle de scène audiovisuelle et d'exhiber les performances de notre règle de décision statistique.

[1] http ://jan.nesvadba.info/cv

L'ensemble de ces tests et des résultats obtenus pour ces travaux sont exposés dans le chapitre 10 de la partie IV.

Troisième partie

Analyse et Indexation des flux audio

Chapitre 7

Détection des silences

Dans ce chapitre, nous montrons les travaux que nous avons menés concernant la détection des silences dans les bandes sonores des flux audiovisuels. Dans la première partie de ce chapitre, nous adressons le problème de la détection des silences. Puis nous présentons les méthodes mises en œuvre.

7.1 Problématique

Dans le domaine de l'indexation cross-média des vidéos numériques, des techniques d'analyse et de segmentation dans les domaines audio et vidéo doivent être développées. Nous nous intéressons ici au domaine audio et plus particulièrement à la détection des silences dans les bandes sonores des flux audiovisuels. En effet, les études que nous avons menées sur des grandes bases de données audiovisuelles ont révélé que plus de la moitié des cas des frontières de scènes étaient caractérisées par la présence de segments de silence dans le flux audio. C'est la raison pour laquelle, nous avons concentré nos travaux sur la détection des silences. De plus, comme nous travaillons avec différents types de contenus nous avons proposé deux méthodes similaires suivant que le signal audio soit compressé ou non.

La partie suivante est consacrée à la présentation en détail de ces deux méthodes.

7.2 Méthode dans les domaines compressé et non compressé

Dans la littérature différents types de détecteurs de silence sont présentés et qui affichent de très bonnes performances [Sou83, LJC87, BCC99]. Le détecteur de silence que nous avons développé est temps réel et compatible à la fois avec les flux compressés et non compressés. Les deux sous-parties suivantes présentent de manière détaillée l'approche utilisée pour chacun des domaines compressé et non compressé. Dans les deux domaines, nos algorithmes reprennent les idées des travaux de [BPDCL02, CSLZ02, Sou83] qui sont les suivantes. Nous utilisons une fenêtre coulissante contenant M fenêtres d'analyse audio de taille L (voir figure 38). Tout d'abord, l'énergie du signal locale est calculée pour chacune des M fenêtres d'analyse audio de taille L de la fenêtre coulissante. Après quoi, la moyenne de l'énergie des M fenêtres est calculée. Enfin, le rapport entre l'énergie moyenne et l'énergie de la fenêtre d'analyse courante

est calculé et comparé à un seuil adaptatif dépendant du niveau de bruit de fond du signal. Si ce rapport (voir équation (193)) est inférieur au seuil alors la fenêtre courante d'analyse est étiquetée comme silence. Pour obtenir une estimation rapide du bruit de fond, nous considérons le minimum de l'énergie calculée du début du signal jusqu'à la fenêtre d'analyse courante.

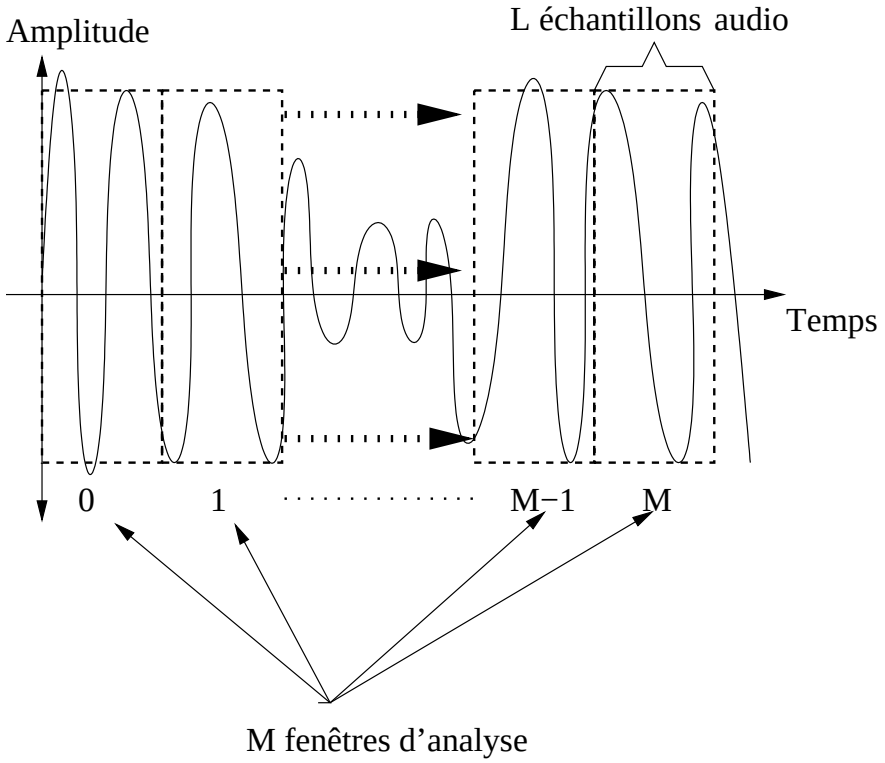


FIG. 38 – Illustration du principe des fenêtres coulissantes

7.2.1 Domaine compressé

La méthode de détection des siences dans le domaine compressé n'a pas été développé dans cette thèse. Nous avons utiliié une méthoe dont Philips Research est propriétaire. Cette méthode est présentée dans l'annexe C.

7.2.2 Domaine non compressé

La méthode utilisée dans le domaine non compressé est identique à celle utilisée dans le domaine compressé. À la différence du domaine non compressé, ici nous travaillons à partir des échantillons audio et de l'énergie du signal et non pas de son spectre. Cette approche se décompose en trois grandes étapes :

1. Calcul de l'énergie du signal (RMS),
2. Mise à jour de la moyenne de la RMS et
3. Détection des possibles frontières de silence.

Étant entendu que nous devons tenir compte des contraintes de temps réel, nous avons développé une approche classique pour son faible coût en termes de temps de calcul.

Enfin, à la différence de la méthode dans le domaine compressé, nous avons entièrement développé la méthode dans le domaine non compressé dans le cadre de cette thèse.

Calcul de l'énergie du signal (RMS)

La première étape consiste donc à évaluer la valeur de l'énergie locale du signal, notée RMS. Cette énergie est calculée pour chacune des fenêtres d'analyse à l'aide de la formule suivante :

$$RMS(K) = \sqrt{\frac{1}{L} \times \sum_{i=0}^{L-1} x(i)^2} \tag{123}$$

avec $x(i)$ le i-ième échantillon audio de la fenêtre d'analyse d'indice K et L représente la taille de la fenêtre d'analyse en terme de nombre d'échantillons.

Mise à jour de la moyenne de la RMS

La mise à jour de la valeur de la moyenne de l'énergie du signal constitue la seconde phase du système. De la même manière que pour le domaine compressé, la moyenne est calculée à partir des valeurs locales de l'énergie des M précédentes fenêtres d'analyse consécutives, incluant la fenêtre K comme étant la dernière de l'ensemble. La valeur de moyenne est donnée par l'équation suivante :

$$RMS_A(K) = \frac{1}{M} \sum_{i=0}^{M-1} log(RMS(K-i)) \tag{124}$$

Cette valeur représente la moyenne logarithmique de l'énergie du signal audio. Nous considérons la moyenne logarithmique pour les mêmes raisons évoquées lors de la sous-partie précédente concernant le domaine compressé.

Détection des possibles frontières de silence

Enfin, la règle de décision utilisée est la suivante. Une fenêtre audio est considérée comme silence si la condition suivante est respectée :

$$\frac{log(RMS(K))}{RMS_A(K)} \leq T_{uncompressed} \tag{125}$$

avec $T_{uncompressed}$ (différent de $T_{compressed}$) un seuil adaptatif pour le domaine non compressé obtenu de manière expérimentale pour chaque document audiovisuel considéré. Nous supposons donc que chaque document a été caractérisé par son propre niveau de bruit. Sur un nombre fixe de fenêtres d'analyse au début d'un document ($N = 10$), nous avons retenu le minimum des valeurs de RMS_i avec $i = 1...N$. Cette valeur est notée RMS_i^*. Le rapport (voir équation (125)) $\frac{RMS_i^*}{RMS_A(K)}$ a été considéré comme le seuil "$T_{uncompressed}$ pour la détection des silences sur le reste du document.

7.3 Résultats et expérimentations

Diverses expérimentations ont été réalisées afin de valider les méthodes proposées pour la détection des silences dans les flux compressés et non compressés. Ces expériences ont été menées sur des bandes sonores extraites de contenus audiovisuels de différents genres.

L'ensemble de ces tests et des résultats obtenus pour ces travaux sont exposés dans le chapitre 10.1 de la partie IV.

Chapitre 8

Détection des transitions Bruit/Bruit dans le flux audio

Dans ce chapitre, nous présentons la méthode de segmentation aveugle proposée pour la détection des transitions de type Bruit/Bruit dans les flux audio continus. La caractérisation de telles transitions est une étape essentielle pour la segmentation en scènes des contenus multimédia par la méthode d'indexation cross-média présentée dans cette thèse. Selon les études menées sur des corpus réels, que nous présentons dans la partie dédiée aux résultats expérimentaux de cette thèse (voir partie IV), les changements de types de bruits occupent une place importante dans l'ensemble des frontières de scènes après les silences. Cette étude s'inscrit donc dans la continuité du cadre général de cette thèse.
Dans une première partie, nous posons le problème de localisation des transitions de bruits dans une bande audio. Puis nous introduisons les types, ou classes, de bruits que nous avons considérés ainsi que les descripteurs audio utilisés pour les caractériser au mieux. Enfin, la dernière partie de ce chapitre est consacrée au développement de la méthode de décision statistique mise en œuvre pour la localisation des discontinuités dans les flux audio.

8.1 Problématique

Les avancées dans le domaine des technologies liées aux télécommunications permettent d'avoir accès à un large choix de contenus audiovisuels. Afin de gérer une telle quantité d'informations des méthodes de gestion automatiques sont nécessaires. Dans les dix dernières années, de nombreux auteurs [KGOG03, CCK+03] ont proposé des méthodes d'indexation des documents audiovisuels basés sur l'analyse et l'extraction de descripteurs bas niveau. Parallèlement à ces travaux, une forte évolution dans les méthodes de traitement du flux audio a été constatée [MB03, SS97]. Les méthodes de traitement des flux audio offrent de grandes perspectives dans différents domaines applicatifs, en particulier pour la classification audio en musique, parole, silence et bruit. Toutefois, la classe audio des bruits ne fait pas l'objet de beaucoup d'attention. Ceci s'explique par la nature complexe de cette classe et l'absence d'organisation à l'intérieur de celle-ci. La classe des bruits est souvent considérée comme la classe par défaut car elle contient tous les sons n'étant ni de la parole, ni de la musique et ni du silence. Enfin, la bande sonore de nombreux contenus audiovisuels de tout genre est composée de segments

de bruits : reportages animaliers, journaux télévisés, rencontres sportives, films, etc ...

Dans ce chapitre de thèse, nous nous intéressons au problème de la caractérisation des sons bruités et plus particulièrement à la détection des transitions de type bruit/bruit dans les bandes sonores. Nous supposons, dans ces travaux, qu'une première classification en parole, musique, bruit et silence a déjà été effectuée et nous nous focalisons sur les segments de bruits détectés. La résolution du problème de la détection des transitions bruit/bruit se décompose en deux phases : sélectionner un ensemble de descripteurs pertinents pour caractériser les sons bruités et proposer une méthode pour caractériser les transitions entre les bruits.

Dans un premier temps, nous nous intéressons aux types de bruits ainsi qu'aux descripteurs permettant leur caractérisation.

8.2 Types de bruits et descripteurs utilisés

8.2.1 Bruits colorés

Dans les modèles spectraux, les sons bruités sont souvent considérés comme des bruits blancs filtrés [SS90]. Cette hypothèse repose sur l'importance de l'enveloppe spectrale dans la perception des sons. En effet, le timbre est corrélé avec l'enveloppe spectrale ainsi qu'avec ses variations dans le temps [SS90].

L'appellation *Bruits Colorés* provient de l'analogie avec les longueurs d'onde de la lumière. Le nom de la couleur qui correspond à une certaine longueur d'onde est affecté au bruit dont le spectre est similaire. Le principal exemple est celui du bruit blanc qui est caractérisé par un spectre uniforme sur l'ensemble des fréquences, ce qui correspond au spectre de la lumière blanche.

Description et exemples

La première catégorie des sons bruités est, intuitivement, la catégorie des sons composés de sons qui peuvent être parfaitement synthétisés par filtrage de bruit blanc : toutes les propriétés perceptives de ces sons sont supposées être contenues dans les enveloppes spectrales à court-terme. Il est important de noter que cette définition est restrictive car, en pratique, peu de sons bruités naturels peuvent être parfaitement synthétisés par filtrage de bruit blanc [HDC04]. D'autres propriétés perceptives importantes doivent être prises en compte, comme par exemple la densité spectrale [HMGB86], l'harmonicité, etc.... Cette catégorie inclut, donc, les sons bruités dont ces propriétés ne sont pas significatives.

Il y a de nombreux exemples de tels sons bruités : flux/reflux de la mer, le vent, etc...

Descripteurs

La caractéristique principale de ces sons est l'évolution de leurs enveloppes spectrales au court du temps. Ces spectres sont obtenus après transformation par la transformée de Fourier rapide (FFT). Néanmoins, il est important de déterminer un ou plusieurs paramètres pour décrire cette propriété.

Les descripteurs les plus couramment utilisés dans la littérature sont :

- Le **roll-off spectral** (cf équation (60) dans la section 3.1.3) est défini dans [SS97]. Il permet d'obtenir la proportion d'énergie dans les hautes fréquences. Il est particulièrement utile pour discriminer les parties voisées des parties non-voisées de la parole. Cependant, l'utilisation de ce descripteur dans le cas des sons bruités n'est pas suffisamment discriminant car deux bruits colorés peuvent avoir deux énergies spectrales différentes mais avoir la même valeur pour le roll-off spectral.
- Le **centroïde spectral** correspond au centre de gravité du spectre d'amplitude, il a été défini par l'équation (59) dans la section 3.1.3. Ce paramètre est comparable au roll-off spectral car il donne une mesure de la proportion d'énergie dans les hautes fréquences du spectre. Ce descripteur n'est cependant pas le plus pertinent pour notre problème pour la même raison que pour le roll-off spectral.
- Le **flux spectral** défini dans [TC02] comme la différence quadratique entre les amplitudes normalisées $\mathcal{S}$ des spectres successifs :

$$F_r = \sum_{k}^{\frac{N}{2}} (\mathcal{S}_r(k) - \mathcal{S}_{r-1}(k))^2 \tag{126}$$

  Ce descripteur est une mesure des variations d'amplitudes du spectre pendant une très faible période de temps (durée de la fenêtre d'analyse). Il pourrait être déterminant pour caractériser les transitions entre les bruits, mais nous ne pouvons pas supposer que ces variations soient des caractéristiques perceptuelles de n'importe quel son bruité.
- La **différence spectrale** est simplement la différence entre les représentations de deux enveloppes spectrales successives. De la même manière que pour le descripteur précédent, la différence spectrale ne permet de caractériser que des variations locales des enveloppes. Cependant, il ne peut être affirmé que ces variations locales sont une caractéristique de perception d'un bruit coloré.

Dans le but de pouvoir discriminer avec plus de précision les bruits colorés dont les enveloppes spectrales sont différentes, nous proposons de définir des descripteurs qui permettent de décrire avec précision les différentes enveloppes spectrales. Dans la littérature, des ouvrages portant sur les domaines de l'analyse et de la synthèse [SS90], de la psychoacoustique [ZF99] ou de la classification de la musique [TC02] introduisent différentes représentations pour l'enveloppe spectrale telles que :

- La première approximation est basée sur les coefficients obtenus par l'application de la méthode du codage linéaire prédictif (LPC). Cette technique est souvent appliquée dans le domaine de la parole. Cependant, il semble ambiguë de comparer le i-ème coefficient de deux fenêtres successives car tous les coefficients ne sont pas corrélés avec les propriétés perceptives des sons. Ceci implique que deux sons peuvent être perçus de manière différente alors que les i-è*mes* coefficients LPC sont très proches.
- Une autre méthode d'approximation des enveloppes spectrales consiste à calculer les coefficients cepstraux (CC). Cette méthode présente les mêmes limitations que la méthode des coefficients LPC. Seuls quelques coefficients peuvent différer pour deux sons perceptiblement différents [ENRC03].
- Dans le but de définir des descripteurs qui sont en relation avec la perception, les coefficients cepstraux peuvent être calculés en utilisant l'échelle Mel [RJ93]. Les coefficients MFC sont souvent utilisés dans le domaine de la parole. Même si ces descripteurs sont

appropriés pour les domaines de la parole et la musique [Log00], leur application aux sons bruités pose le même problème que pour les méthodes précédentes (LPC et CC).

Nous proposons ici d'appliquer les résultats de la recherche de Goodwin concernant la modélisation des sons [Goo96]. Ces travaux sont liés au modèle de perception pour les bruits selon lequel un bruit est représenté correctement par les variations temporelles dans les Bandes Rectangulaires Équivalentes (ERBs). Cependant, nous proposons d'adapter cette méthode en utilisant l'échelle Barks plutôt que les ERBs. Nous avons fait ce choix car le nombre de bandes de Barks est inférieur à celui des ERBs [SA99] : ce choix implique une réduction du nombre de descripteurs. L'échelle Barks est une échelle psychoacoustique introduite par Zwicker [ZF99]. L'échelle est composée de 26 sous-bandes incluant l'ensemble des fréquences audibles, de 0 à 22050Hz, la moitié du taux d'échantillonnage couramment utilisé $R = 44100$Hz.

Les enveloppes spectrales à court-terme des sons bruités sont représentées par les proportions d'énergie dans chaque sous-bande de Barks :

$$\forall f, \quad S'(f) = \sqrt{\frac{1}{B_k} \sum_{f \in \text{Bande de Barks } k} S^2(f)} \tag{127}$$

avec B_k la taille de la bande de Barks d'indice k et S le spectre discret à court-terme.
Ainsi, 26 descripteurs caractérisent la couleur des sons bruités. Les variations de quelques uns, des coefficients, sont suffisantes pour signifier un changement de couleur du son. Ce changement peut-être perçu du fait que chaque descripteur est lié à la perception. La figure 39 montre un exemple de représentation des bandes de Barks pour le cas d'un son d'écoulement de l'eau d'un torrent. L'axe des abscisses représente l'indice de la bande de Barks, l'axe des ordonnées est l'indice de la fenêtre d'analyse et le dernier axe caractérise la valeur de la proportion d'énergie comprise entre 0 et 1.

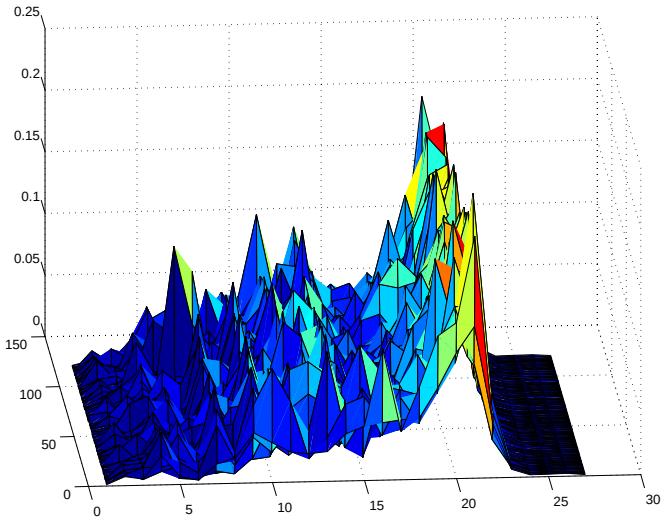


FIG. 39 – Exemple de représentation des bandes de Barks pour un son de torrent.

8.2.2 Bruits impulsifs

Les recherches en psychoacoustique, sur la perception des bruits, ont montré que peu de propriétés audibles permettaient de discriminer cet ensemble de bruits [Har97]. Une de ces rares propriétés est la variation brusque de l'intensité. Nous proposons, donc, une classe des sons bruités naturels qui caractérisent ces pulsations : les sons composés d'impulsions régulières ou irrégulières.

Description et exemples

En effet, de nombreux sons naturels bruités sont composés d'impulsions périodiques ou non-périodiques. Par exemple, les bruits d'applaudissements, de pas humains, de la pluie qui tombe, etc ... Nous appelons ces bruits, les *bruits impulsifs* [Har97]. La fréquence des pulsations composant ce type de sons doit être inférieure à environ 20Hz. Sinon cette fréquence est détectée par le système auditif humain comme une hauteur.

Descripteurs pour les bruits impulsifs

De nombreux modèles d'analyse/synthèse de sons dissocient les transitoires des sinusoïdes et des parties stochastiques du signal [VM00]. C'est pourquoi, de nombreuses méthodes pour la détection des transitoires basées sur les variations de l'énergie [Lev98], sur les variations de l'énergie des hautes fréquences [MB96] ou sur les variations de l'amplitude du spectre [PR99] ont été développées. Cependant, nous pensons que ces méthodes ne peuvent pas être de bonnes candidates pour la caractérisation des bruits impulsifs. En effet, la présence d'énergie dans les hautes fréquences peut indiquer la présence d'impulsions mais simplement aussi le niveau de bruit. Ainsi, la définition d'un autre descripteur spécifique aux sons impulsifs est nécessaire.

C'est pourquoi, nous voulons choisir un descripteur audio permettant de caractériser les sons bruités impulsifs. Les sons impulsifs sont composés de pulsations. La première étape consiste à détecter les fenêtres d'analyse qui contiennent une pulsation perçue. De nombreux descripteurs audio peuvent être utilisés pour caractériser ces fenêtres, comme par exemple le taux de passage par zéro (ZCR) ou bien les variations de l'énergie. Cependant, nous avons choisi d'utiliser une autre approche, plus en adéquation avec le contexte des sons bruités [Han03]. Cette méthode est basée sur la mesure du kurtosis (moment statistique d'ordre 4, voir équation (58) dans la section 3.1.3). Une forte valeur du kurtosis (supérieure à 10) signifie que nous sommes en présence d'une fenêtre d'analyse de type impulsif. Dans le cas contraire, nous considérons que cette fenêtre d'analyse n'est composée d'aucune pulsation.

Une valeur de kurtosis est affectée à chaque fenêtre. Un seuil est alors fixé dans le but de définir les fenêtres qui sont considérées comme impulsives. Il est important de noter que ce seuil est indépendant du niveau de volume du son analysé. Plus l'impulsion est forte et audible, plus la valeur du kurtosis associée est élevée. Cependant, la valeur du kurtosis n'est pas seulement un indicateur de la présence d'impulsion, mais elle est utilisée pour caractériser la nature de l'impulsion.

Nous pensons qu'il est aussi important d'être capable de discriminer les sons bruités impulsifs qui ne diffèrent pas par la nature des pulsations mais par leur périodicité. C'est la raison pour laquelle nous proposons d'utiliser la fréquence des pulsations en plus du kurtosis pour

caractériser les sons impulsifs. Après avoir calculé le kurtosis sur l'ensemble du signal, nous marquons les fenêtres dont la valeur du kurtosis excède un seuil de tolérance fixé de manière expérimental. Après quoi, l'écart temporel entre deux fenêtres marquées est estimé, permettant ainsi de calculer la fréquence des impulsions. Nous travaillons ainsi avec une longueur d'analyse équivalent à la durée de l'échantillon étudié. Dans le cas de la détection d'une seule pulsation, la fréquence est nulle car pour estimer la fréquence des impulsions, au moins deux pics dans la courbe des valeurs du kurtosis sont nécessaires. Nous estimons que si une seule pulsation a été détectée il peut s'agir d'une imprécision de calcul, donc nous ne la prenons pas en considération.

Par conséquent, toutes les fenêtres analysées sont caractérisées par un couple de valeurs : le kurtosis et sa fréquence. Le kurtosis permet de discriminer les sons colorés des sons impulsifs, tandis que la périodicité est utilisée pour discriminer les sons impulsifs entre eux. Le kurtosis semble être un bon candidat pour l'estimation de la présence de transitoires dans les sons bruités [Han03]. Enfin, la taille de la fenêtre d'analyse est un paramètre important car elle définit l'hypothèse selon laquelle le signal du son est stationnaire [GPD00].

Nous montrons dans la figure 40 les variations du kurtosis en fonction du temps pour un son de machine à écrire. L'axe des abscisses représente l'index des fenêtres d'analyse alors que l'axe des ordonnées correspond à la valeur du kurtosis. Dans ce travail, nous avons utilisé des fenêtres d'analyse comprenant 512 échantillons avec une fréqence d'échantillonnage de 44100 Hz. Cette taille de fenêtre est la plus couramment utilisée dans la littérature. De plus, nous avons fixé un seuil pour le kurtosis (ligne horizontale sur la figure) égal à 5. Les expérimentations menées dans [Han03] ont montré qu'un seuil supérieur à 4 est suffisamment descriminatif pour caractériser une impulsion à partir de la valeur du kurtosis.

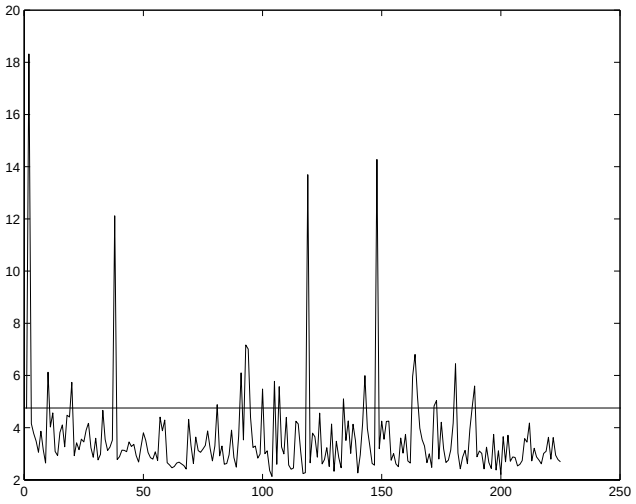


FIG. 40 – Exemple de valeurs du kurtosis pour un son de machine à écrire avec une fenêtre d'analyse de taille 512 échantillons.

8.2.3 Bruits pseudo-périodiques

Dans cette sous-partie, nous présentons ce qui caractérise les bruits pseudo-périodiques et ce qui les différencie des autres types de bruits.

Description et exemples

De nombreux sons naturels bruités sont caractérisés par la présence d'une hauteur qui peut-être perçue par le système auditif humain. Parmi ces sons, il y a par exemple les bruits des machines, les insectes volants, etc ...
Une hauteur peut être perçue avec des amplitudes différentes. De plus, différentes hauteurs peuvent être perçues en même temps. Ce phénomène peut être expliqué par les raisons suivantes :

1. Les bruits peuvent être modélisés comme une somme de sinusoïdes d'après le modèle de bruit thermal [Har97]. La fréquence des sinusoïdes est supposée être espacée de manière constante. Les expérimentations concernant la modélisation spectrale des bruits à l'aide de sinusoïdes à court-terme ont montré que de choisir un faible nombre de sinusoïdes implique la perception d'une hauteur dans ces bruits [HBDC02]. Ces expérimentations ont aussi montré que certains sons bruités naturels pouvaient être considérés comme harmoniques ou pseudo-harmoniques [Han03].
2. La perception de la hauteur est à l'origine d'études dans le domaine de la psychoacoustique. Un type de son synthétique est particulièrement expérimenté : les bruits ondulants (rippled noises). Ce type de bruit est synthétisé par l'ajout à un bruit blanc d'une version retardée et atténuée de lui-même [Yos96]. Les sons ondulants (rippled noises) laissent percevoir une hauteur de par la périodicité du signal.

Descripteur pour les bruits pseudo-périodiques

Ce descripteur que nous définissons doit permettre de décrire avec la plus grande précision possible la hauteur d'un son selon deux caractéristiques : l'amplitude de la hauteur et sa valeur. Quelques méthodes ont été proposées dans le domaine de la psychoacoustique pour mesurer l'amplitude d'une hauteur pour un son ondulant [Yos96]. La méthode que nous avons choisie considère la fonction d'autocorrélation, Γ, et plus précisément le rapport d'autocorrélation du second maximum de la fonction d'autocorrélation, noté $\Gamma(\tau)$, et sa première valeur que l'on note $\Gamma(0)$. Cette première valeur de la fonction d'autocorrélation représente l'énergie totale du signal. Ce rapport d'autocorrélation, AR, est défini par l'équation suivante :

$$AR = \frac{\Gamma(\tau)}{\Gamma(0)} \tag{128}$$

Ce descripteur est aussi un bon candidat pour les problèmes de segmentation et de reconnaissance dans le domaine de la parole [ZSN03] et d'estimation de la hauteur des sons harmoniques. Néanmoins, au regard des expériences que nous avons menées, le rapport d'autocorrélation semble être un bon candidat pour caractériser les sons bruités de type pseudo-périodiques.

En outre, deux bruits pseudo-périodiques peuvent être différents et avoir le même rapport d'autocorrélation. Dans ce cas, la différence se situe au niveau des hauteurs perçues. Les deux

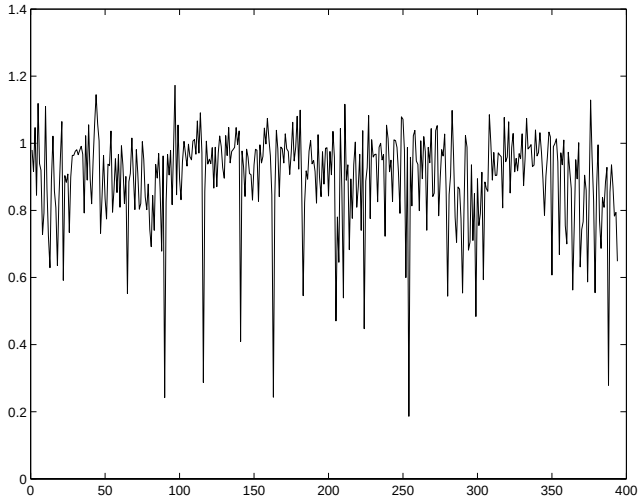


FIG. 41 – Exemple de valeurs de l'AR pour un son de rasoir électrique.

sons ne laissent pas percevoir la même hauteur et pourtant le rapport d'autocorrélation obtenu pour ces deux sons est identique. C'est pourquoi, nous proposons un deuxième descripteur pour caractériser au mieux les bruits pseudo-périodiques. Il s'agit de la période du son analysé, p, définie par :

$$p = \frac{\tau}{R} \tag{129}$$

avec τ l'indice du second maximum de la fonction d'autocorrélation et R la fréquence d'échantillonnage.

La figure 41 illustre les variations du rapport d'autocorrélation en fonction du temps dans le cas d'un bruit provenant d'un rasoir électrique. L'axe des abscisses représente l'index des fenêtres d'analyse alors que l'axe des ordonnées correspond à la valeur du rapport d'autocorrélation. Comme il est possible de le remarquer sur la figure 41, certaines valeurs du rapport d'autocorrélation peuvent prendre des valeurs supérieures à 1. Ceci s'explique par des imprécisions de calculs liées à la propagation d'erreurs dans le calcul des valeurs de la fonction d'autocorrélation.

8.2.4 Bruits avec sinusoïdes

Dans cette sous-partie, nous présentons ce qui caractérise les bruits composés de sinusoïdes et ce qui les différencient des autres types de bruits.

Description et exemples

Les sons réels sont le plus souvent supposés comme étant un mélange de plusieurs sources, harmoniques ou non. Si une ou plusieurs de ces sources est harmonique (ou pseudo-harmonique) et si leur niveau sonore n'est pas trop fort, ces sources peuvent être perçues : elles représentent donc une caractéristique perceptive importante pour cet ensemble de sons.

Les exemples de sons bruités réels qui sont composés de sinusoïdes sont nombreux. En particulier, il y a : le vent dans les arbres avec des oiseaux qui chantent, le bruit des embouteillages

dans la rue, etc...

Descripteurs pour les bruits avec sinus

Les bruits naturels sont représentés par des spectres d'amplitude à court terme qui sont composés de pics qui correspondent aux sinusoïdes contenues dans ce son. Le descripteur associé à ce type de bruit est naturellement le nombre de sinusoïdes. De nombreuses méthodes ont été proposées dans le contexte de l'analyse et de la synthèse de sons :

- La plus classique des méthodes consiste à considérer tous les maxima locaux des spectres d'amplitudes comme étant des sinusoïdes. Cette méthode a été développée dans [SS90]. Cependant, la précision de cette méthode est directement liée au niveau de volume du son analysé. De ce fait, nous ne pouvons pas utiliser cette méthode dans notre problème.
- Une autre méthode consiste à caractériser les sinusoïdes comme la différence entre les pics et l'amplitude de ses deux voisins. Ici aussi, le niveau de volume du son analysé est à prendre en considération. Cette méthode n'est donc pas applicable.
- Une approche originale pour la caractérisation des sinusoïdes a été proposée dans [PR98]. Cependant, cette méthode n'apparaît pas être suffisamment robuste dans le cas de sons bruités car des expériences menées sur des sons synthétiques montrent que cette méthode fournit une sur-estimation du nombre de sinusoïdes réellement présent dans le signal étudié [Han03]
- Une nouvelle méthode a été proposée et expérimentée comme étant plus précise lorsqu'il s'agit de sons bruités [Han03, HDC03]. Cette méthode est basée sur une analyse statistique de l'intensité des fluctuations de l'enveloppe temporelle pour une bande fréquentielle étroite du signal étudié. Si cette bande n'est composée que de bruit, les fluctuations sont importantes. Si cette bande est composée d'une sinusoïde (même noyée dans du bruit), les fluctuations sont alors plus faibles. Par conséquent, cette approche propose d'analyser les pics du spectre à court terme et enfin d'estimer quels sont les pics qui correspondent réellement à des sinusoïdes et ceux qui sont du bruit ou des artefacts. Une mesure pour chaque casier de l'amplitude du spectre est effectuée et un seuil défini au préalable permet de définir le nombre de casiers qui correspondent aux sinusoïdes. Ainsi, un nombre de sinusoïdes est assigné à chaque fenêtre d'analyse. Donc, cette dernière méthode nous semble être une bonne candidate pour estimer le nombre de sinusoïdes d'un son bruité.

Le tableau de la figure 42 montre le nombre moyen de sinusoïdes détectées dans différents types de sons bruités naturels. Nous pouvons remarquer que ce nombre varie fortement suivant les types de sons bruités considérés : vent dans les feuilles (bruit coloré), saxophone (bruit harmonique) ou extrait d'un match de football (mélange de bruits).

Son	Nombre de sinusoïdes
Vent dans les feuilles	61.26
Saxophone	224.95
Match de Football	54.63

Fig. 42 – Résultats de l'estimation du nombre de sinusoïdes.

Enfin, le tableau de la figure 43 résume les descripteurs utilisés en fonction du type de bruit.

Type de bruit	**Descripteurs**
Coloré	Énergies dans les bandes de Barks
Pseudo-périodique	Rapport d'autocorrélation Période du son bruité
Avec sinus	Nombre de sinusoïdes
Impulsif	Kurtosis Fréquence des impulsions

FIG. 43 – Type de bruit et descripteurs associés.

8.3 Segmentation des bruits

Cette section est dédiée à la présentation détaillée de la méthode de décision statistique utilisée pour caractériser les transitions entre les différents types de bruits que nous considérons.

Le schéma de détection des transitions audio est basé sur le principe du modèle de décision statistique bayésien. Il s'agit d'une méthode de décision classique et régulièrement utilisée dans la littérature depuis plus de 15 ans [CH78] et encore de nos jours [Del00, Mor04]. Nous réutilisons le schéma de décision bayésienne avec deux hypothèses dans le cas de lois Normales multidimensionnelles. Nous avons déjà présenté ce schéma dans le cas mono-dimensionnel pour la détection des frontières de scènes (voir section 6.1).

Ici, pour notre problème, la variable stochastique, X_i, représente le vecteur des descripteurs audio extraits au cours du temps, i, correspondant à l'événement X. Ce vecteur est de la forme :

$$X_i = \begin{pmatrix} x_{1i} \\ . \\ . \\ . \\ x_{mi} \end{pmatrix} \tag{130}$$

avec m le nombre total de descripteurs audio utilisés.

La variable, h, est distribuée sur l'ensemble des hypothèses. Dans notre cas, nous considérons uniquement deux hypothèses, H_1 et H_2 :

- H_1 : Absence de transition audio au temps t_0 et
- H_2 : Présence d'une transition audio au temps t_0.

Dans la suite, nous supposons que les éléments du vecteur de descripteurs, X_i, suivent la distribution Normale. Ainsi, nous allons considérer une distribution gaussienne multidimensionnelle pour X_i à l'intérieur de chaque segment audio homogène compris dans l'intervalle $[t_0 - n, t_0 + n]$ où n est un paramètre entier. De ce fait, une transition dans le flux audio au temps t_0 peut-être exprimée à l'aide de deux hypothèses comme suit :

- $H_1 : \forall t, t_0 - n \leq t \leq t_0 + n$, la variable multidimensionnelle, X, suit la même distribution normale définie par $N_0(\mu_0, \Sigma_0)$ avec μ_0 le vecteur moyenne et Σ_0 la matrice de covariance.
- $H_2 : \forall t, t_0 - n \leq t \leq t_0$, X suit la distribution normale définie par $N_1(\mu_1, \Sigma_1)$ avec μ_1 le vecteur moyenne et Σ_1 la matrice de covariance. Et $\forall t, t_0 \leq t \leq t_0 + n$, X suit la distribution normale définie par $N_2(\mu_2, \Sigma_2)$avec μ_2 le vecteur moyenne et Σ_2 la matrice de covariance.

L'hypothèse H_2 correspond à une transition de bruit, et H_1 implique la continuité du signal audio. Considérons la définition précédente pour les hypothèses H_1 et H_2, le rapport de vraisemblance s'exprime de la manière suivante :

$$\frac{L1}{L2} = \frac{A}{B} \tag{131}$$

avec $A = \prod_{t=t_0-n}^{t_0+n} \frac{1}{(2\pi)^{\frac{m}{2}} \times \sqrt{det(\Sigma_0)}} \times e^{-\frac{1}{2}(X_t^T \times \Sigma_0^{-1} \times X_t)}$ et

$$B = \prod_{t=t_0-n}^{t_0-1} \frac{1}{(2\pi)^{\frac{m}{2}} \times \sqrt{det(\Sigma_1)}} \times e^{-\frac{1}{2}(X_t^T \times \Sigma_1^{-1} \times X_t)} \times \prod_{t=t_0}^{t_0+n} \frac{1}{(2\pi)^{\frac{m}{2}} \times \sqrt{det(\Sigma_2)}} \times e^{-\frac{1}{2}(X_t^T \times \Sigma_2^{-1} \times X_t)}$$

avec Σ_z la matrice de covariance, $z = 0, 1, 2$.
Si X_i est une suite finie de variables aléatoires, alors la matrice de covariance des X_i est la matrice carrée dont le coefficient en (i, j) est donné par :

$$c_{i,j} = \mathrm{Cov}(X_i, X_j)$$

avec $Cov(A, B)$ la covariance des variables aléatoires A et B.
Une matrice de covariance est toujours symétrique. Dans notre cas, la matrice de covariance, Σ, peut-être définie de la manière suivante [And03] :

$$\Sigma = \begin{pmatrix} E((x_{1k} - E(x_{1k})) \times (x_{1k} - E(x_{1k}))) & \cdots & E((x_{1k} - E(x_{1k})) \times (x_{mk} - E(x_{mk}))) \\ \vdots & \ddots & \vdots \\ E((x_{mk} - E(x_{mk})) \times (x_{1k} - E(x_{1k}))) & \cdots & E((x_{mk} - E(x_{mk})) \times (x_{mk} - E(x_{mk}))) \end{pmatrix}$$

avec E l'espérance mathématique et x_{ki} les k composantes du vecteur de descripteurs X_i.
Le schéma de la figure 44 illustre la définition que nous avons formulée pour les hypothèses H_1 et H_2.

Suivant le développement classique [Ven00] qui consiste à considérer le logarithme de rapport de vraisemblance (voir équation 131) et à comparer le tout à un seuil probabiliste, nous avons :

$$M = \frac{n}{2}\left[ln(det(\Sigma_1)) + ln(det(\Sigma_2))\right] - n \times ln(det(\Sigma_0) + \frac{1}{2}[C + D - E] \tag{132}$$

avec $C = \sum_{t=t_0-n}^{t_0-1} \Sigma_1^{-1} \times X_t)$, $D = \sum_{t=t_0}^{t_0+n} (X_t^T \times \Sigma_2^{-1} \times X_t)$ et $E = \sum_{t=t_0-n}^{t_0+n} (X_t^T \times \Sigma_0^{-1} \times X_t)$
et :

$$M \begin{matrix} < \\ > \end{matrix} 2 \times n \times ln(l \times P/(1-P)) \begin{matrix} => H2 \\ => H1 \end{matrix} \tag{133}$$

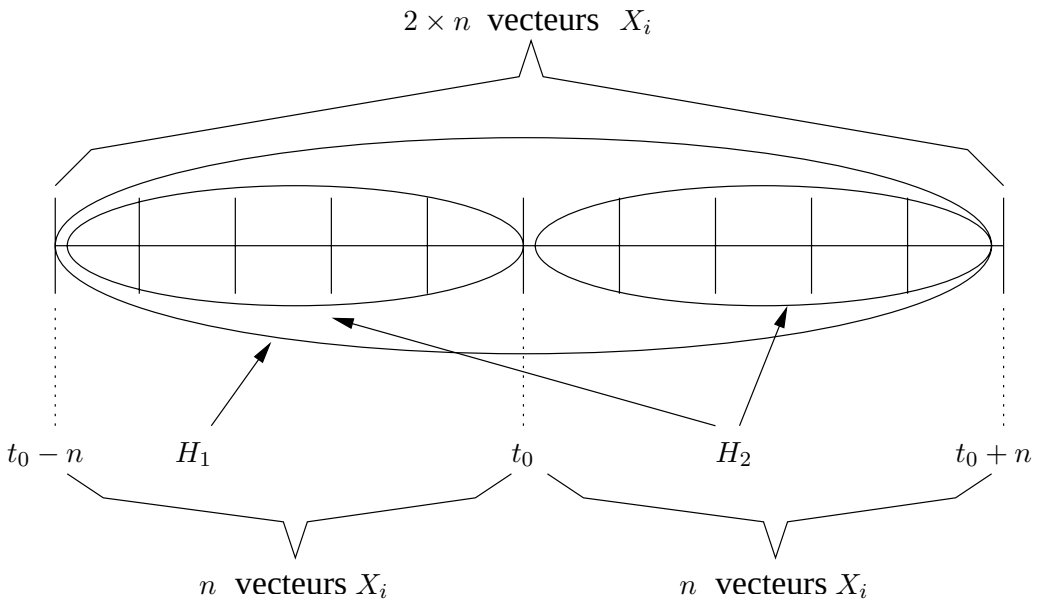


FIG. 44 – Illustration de la définition des hypothèses H_1 et H_2

où P est la probabilité relative à l'hypothèse $H2$ - présence d'une transition de bruit.

Dans le cas où les matrices de covariances, Σ_j avec $j = 1, 2, 3$, ont été estimées à partir des mêmes échantillons contenus dans le vecteur X_i et si, selon notre hypothèse, les composantes du vecteur X_i sont indépendantes entre elles, alors la somme $A + B - C$ de l'équation (132) est nulle. L'expression de M dans l'équation (132) peut être ré-écrite de la manière suivante [And03] :

$$M = \frac{n}{2}\left[ln(det(\Sigma_1)) + ln(det(\Sigma_2))\right] - n \times ln(det(\Sigma_0) \tag{134}$$

Enfin, considérons l'équation (133) avec $l = 1$ et l'expression de M simplifiée donnée par l'équation (134), nous obtenons donc notre règle de décision finale :

$$M \begin{array}{c} < \\ > \end{array} 2 \times n \times ln(P/(1-P)) \begin{array}{l} => H2 \\ => H1 \end{array} \tag{135}$$

Pour résumer, le principe de notre méthode de segmentation consiste à extraire un vecteur, X_i, de descripteurs dans $2 \times n$ fenêtres d'analyse successives. Puis, nous posons deux hypothèses selon lesquelles soit l'ensemble des $2 \times n$ vecteurs de descripteurs suivent la même distribution normale avec les mêmes paramètres (équivalent à H_1) soit les n premiers vecteurs suivent une distribution et les n suivants une distribution normale avec des paramètres différents (équivalent à H_2). La méthode de décision (voir équation 133) est ensuite appliquée pour caractériser l'hypothèse la plus probable entre H_1 et H_2. La mise en œuvre de cet algorithme nécessite l'estimation des valeurs pour la probabilité P et pour le nombre de fenêtres n. Pour cela, nous avons fait différents tests expérimentaux desquels nous avons pu évaluer les valeurs optimales pour ces paramètres : avec $n = 50$ et $P = 0.1$ notre méthodes offre les meilleures performances.

Enfin, cette démarche est connue dans la littérature sous le nom de segmentation *aveugle*

ou *à priori* dans le sens où il s'agit de la détection d'une transition et non pas de l'identification d'un bruit comme appartenant à une classe connue. Néanmoins, dans certains cas, la connaissance de la classe du bruit obtenue à l'aide d'un processus de classification ou bien connue à priori, peut nous guider dans le choix des descripteurs. Dans ce cas, il s'agit de la détection des transitions entre les bruits de la même classe (pseudo-périodique, impulsif, etc...) opérant sur les descripteurs qui caractérisent cette classe au mieux. Ainsi, nous pouvons qualifier cette segmentation de *semi-aveugle*. Par conséquent, il est judicieux de présenter la classification des bruits que nous avons mise en œuvre.

8.4 Résultats et expérimentations

Diverses expérimentations ont été réalisées afin d'une part de valider l'ensemble des descripteurs proposés et d'autre part, de mettre en avant les poins forts et les points faibles de la méthode de segmentation statistique aveugle. L'ensemble de ces tests et des résultats obtenus pour ces travaux sont exposés dans le chapitre 11 de la partie IV.

Chapitre 9

Classification des bruits

9.1 Problématique

Parmi l'ensemble des tâches relatives à l'indexation multimédia, la classification des bandes sonores est devenue une nécessité. En effet, la tâche de segmentation en scène audio-visuelles ne peut être réalisée sans une méthode d'analyse des flux audio et vidéo en amont. Jusqu'à présent, différentes approches pour la structuration des documents audio-visuels ont été présentées dans [CCK+03, KGOG03]. Ces méthodes consistent à extraire et analyser des descripteurs audio-visuels. En particulier, les algorithmes avancés dans le domaine de la classification audio en musique, parole, silence et bruits ont été développées dans [SS97, MB03]. Les domaines concernant la musique et la parole ont déjà été largement explorés, à savoir la classification des genres musicaux [TC02] ou la reconnaissance automatique de la parole [ZSN03]. En revanche, très peu de travaux ont été proposés, dans la littérature, concernant le traitement des sons bruités, en particulier de la classification [EMSK99].

Restant dans le cadre d'une approche bayésienne, nous proposons donc dans ce chapitre, une méthode statistique pour la classification des sons bruités. Un grand nombre des contenus audio-visuels tels que les documentaires ou les scènes du cinéma muet contiennent des parties sonores bruitées qu'il n'est pas possible de considérer comme de la musique ou de la parole. La segmentation et l'indexation de telles parties requièrent des techniques spécifiques. C'est la raison pour laquelle, la classification des sons naturels bruités est une tâche essentielle dans le cadre de l'indexation multimédia.

Dans la première partie de ce chapitre, nous introduisons les classes de sons naturels bruités que nous allons traiter ainsi que les descripteurs associés. Après quoi, nous présentons le principe de classification statistique mise au point.

9.2 Classes de bruits et descripteurs

Dans [HLDCBP04], un ensemble de descripteurs audio adapté à la segmentation des sons bruités a été proposé. À partir de ces descripteurs, nous définissons six classes de sons naturels bruités et nous réduirons la taille de l'ensemble des descripteurs, utilisés pour la classification statistique, à trois éléments.

Dans cette partie, nous donnons une définition des six classes que nous proposons. Ces six classes sont issues de l'union de trois groupes et de leurs intersections. Les trois groupes que nous proposons sont les suivants : pseudo-périodiques, impulsifs et avec sinusoïdes. Une définition de ces trois types de bruits a déjà été donnée dans la partie 8.2. Toutefois, nous rappelons brièvement la composition de ces trois groupes. Après quoi, nous présentons les six classes proposées. La figure 46 illustre les intersections et les inclusions de ces classes.

Bruits purs

Dans le chapitre précédent (chapitre 8), nous avons introduit les classes de bruits purs, à savoir : impulsifs, pseudo-périodiques et composés de sinusoïdes. Nous avons également proposé les descripteurs qui caractérisent ces classes au mieux.

Dans la suite de ce chapitre, nous dénotons ces classes par groupes G_p pour pseudo-périodique, G_i pour impulsif et G_s pour composé de sinusoïdes respectivement.

Intersection de groupes

Les trois principaux groupes que nous avons définis ne sont pas disjoints. Nous enrichissons, donc, notre classification en définissant les six classes de bruits suivantes. Les intersections de ces groupes G_p, G_i et G_s induisent la définition des six classes que nous proposons. Nous notons G l'ensemble des sons bruités.

- **Sons bruités composés de sinusoïdes** : Comme il a été précédemment expliqué, les sons bruités composés de sinusoïdes sont assimilés aux bruits pseudo-périodiques. Ainsi, nous pouvons dire que :
$$C_s = G_s \subset G_p \tag{136}$$
avec C_s la classe pour les sons bruités composés de sinusoïdes.
- **Les sons bruités pseudo-périodiques et impulsifs** : ces sons laissent percevoir une hauteur. Cette hauteur provient des impulsions périodiques dont les fréquences sont supérieures à 20 Hertz. En d'autres termes, les sons bruités assimilés comme étant pseudo-périodiques et impulsifs sont des sons composés d'impulsions qui peuvent être périodiques. Mais cette périodicité ne peut pas être perçue par le système auditif humain comme une hauteur. Par exemple, les sons d'applaudissements sont considérés comme purement impulsifs alors que les sons de rasoir électrique sont assimilés à la fois impulsifs et pseudo-périodiques (voir la figure 45). En notant C_{ip} la classe des sons bruités naturels à la fois pseudo-périodiques et impulsifs, nous pouvons écrire la relation suivante :
$$C_{ip} = (G_p \setminus C_s) \cap G_i \tag{137}$$
- **Sons bruités pseudo-périodiques** : nous considérons la classe des sons bruités pseudo-périodiques comme l'ensemble des sons purement pseudo-périodiques, c'est-à-dire qui ne sont ni impulsifs ni composés de sinusoïdes. Nous notons C_p cette classes de sons bruités, nous pouvons écrire aussi :
$$C_p = G_p \setminus (C_{ip} \cup C_s) \tag{138}$$

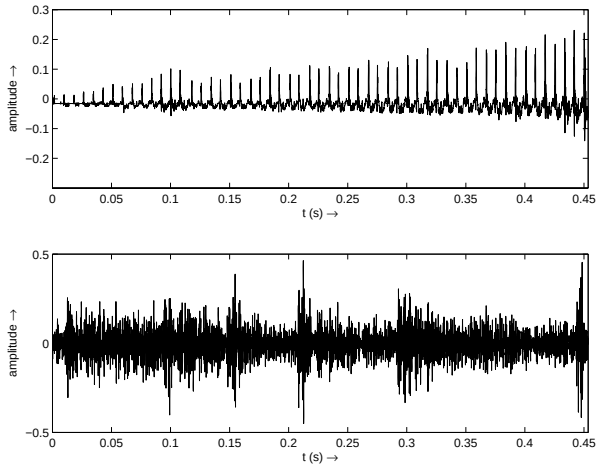


Fig. 45 – Représentation temporelle pour deux sons bruités impulsifs : une hauteur est perceptible pour le premier (rasoir électrique) tandis que la présence d'un rythme caractérise le second (applaudissements).

- **Sons bruités impulsifs composés de sinusoïdes** : ces sons bruités sont composés d'impulsions (périodiques ou non) et de sinusoïdes. Les principaux exemples de sons naturels pour illustrer cette classe sont les sonneries de téléphone, les bruits carillon, etc.... Nous notons C_{is} cette classe, et nous pouvons écrire la relation suivante :

$$C_{is} = C_s \cap G_i \tag{139}$$

- **Sons bruités impulsifs** : nous considérons cette classe comme l'ensemble des sons qui sont considérés comme impulsifs mais qui ne sont ni périodiques (ou pseudo-périodiques), ni composés de sinusoïdes. Cette classe sera notée C_i et nous pouvons aussi dire que :

$$C_i = G_i \setminus (C_{ip} \cup C_{is}) \tag{140}$$

- **Sons bruités colorés** : Cette classe que nous proposons est composée de tous les sons bruités qui ne sont composés d'aucune pulsation (ou rythme) et d'aucune sinusoïde (aucune hauteur n'est perceptible). Par conséquent, cette classe, que nous notons C_c, regroupe l'ensemble des sons qui n'appartiennent à aucune des autres classes précédemment définies. Enfin, nous avons :

$$C_c = G \setminus (C_{ip} \cup C_i \cup C_s \cup C_p \cup C_{is}) \tag{141}$$

La figure 46 illustre les définitions énoncées ci-dessus. Dans, cette figure la notation "$A \rightarrow B$" signifie "A est un sous-ensemble de B".

Algorithme et descripteurs

Dans cette sous-partie, nous présentons un ensemble de descripteurs audio que nous avons choisi d'incorporer dans le schéma de classification statistique que nous proposons. Certains de ces descripteurs ont déjà été présentés dans le chapitre 8 précédent. Néanmoins, dans le

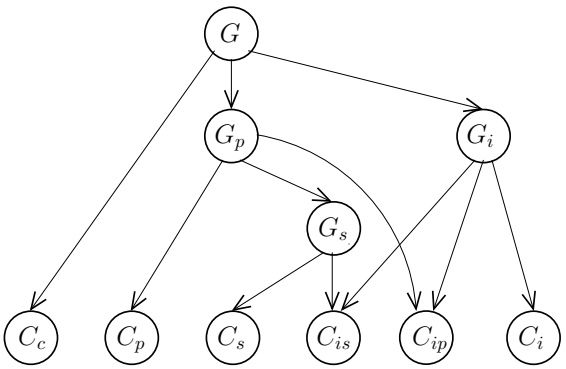


FIG. 46 – Illustration des 6 ensembles proposés pour la classification des sons bruités naturels.

cadre des travaux présentés dans ce chapitre, nous faisons appel à des variantes pour certains de ces descripteurs.
Dans un souci de réduire au maximum la taille de l'espace des descripteurs, nous avons sélectionné uniquement un seul descripteur par classe. Les descripteurs que nous utilisons ne sont pas originaux à proprement parlé, mais l'aspect novateur réside dans l'interprétation que nous en faisons, à savoir, la classification des sons bruités. Les justifications des choix de ces descripteurs utilisés sont données ci-dessous. Ces justifications sont principalement basées sur des recherches expérimentales concernant la perception des caractéristiques des sons.

Fréquence des impulsions

La première étape permet de caractériser les fenêtres d'analyse qui contiennent des impulsions. Avec cette approche, un son bruité impulsif est composé de quelques fenêtres étiquetées impulsives (uniquement lorsqu'une impulsion se produit). Cependant, pour qu'un son bruité soit considéré comme impulsif il faut qu'une grande majorité des fenêtres d'analyse qui le composent soit étiquetée comme impulsive. C'est la raison pour laquelle, nous avons décidé, pour caractériser les bruits impulsifs, de considérer non pas les valeurs du kurtosis mais plutôt la fréquence des impulsions et de proposer l'algorithme suivant pour l'estimer.

Entre deux fenêtres contenant une impulsion (notées n_i et n_{i+1}), toutes les fenêtres sont étiquetées par la valeur de la fréquence F. La valeur de la fréquence dépend à la fois du taux d'échantillonnage R, et de la taille N, de chacune des fenêtres d'analyse (en nombres d'échantillons). La fréquence F, exprimée en Hertz, est donnée par la formule suivante :

$$F = \frac{R}{N(n_{i+1} - n_i)} \tag{142}$$

Il est important de noter que le choix de la fréquence plutôt que de la périodicité est justifiée par la continuité de la fréquence : les sons qui ne sont pas composés d'impulsions ont une fréquence de pulsation nulle, de même pour les sons pour lesquels une seule impulsion a été détectée.

Un autre point important concerne la précision nécessaire pour le calcul de la fréquence des impulsions. Cette précision dépend de la taille des fenêtres d'analyse. Cette taille impose généralement l'utilisation d'une analyse de type *overlap-and-add* dans le but d'améliorer la précision de calcul.

Les sons bruités appartenant à la classe des sons bruités impulsifs sont caractérisés par une fréquence des impulsions élevée. En théorie, il n'y a pas de limites de valeurs pour cette fréquence. Néanmoins, des bornes inférieures et supérieures ont été fixées de manière expérimentale. La borne supérieure peut-être supérieure à la limite audible (20 Hz). Dans ce cas, il s'agit d'un son appartenant à la fois aux classes des sons bruités impulsifs et pseudo-périodiques. En ce qui concerne la borne inférieure, elle est bien évidemment égale à 0.

La force de la hauteur

Dans la partie 8.2, nous avons déjà traité du descripteur utilisé pour caractériser la force de la hauteur. Il s'agit du rapport d'autocorrélation.

Le rapport d'autocorrélation (ACR) est théoriquement compris dans l'intervalle $[0;1]$. En pratique, la propagation des imprécisions de calcul des valeurs de la fonction d'autocorrélation peut conduire à obtenir un rapport légèrement supérieur à 1. Si le rapport est proche de 1, cela signifie qu'une hauteur est perçue. Au contraire, un rapport nul (égal à 0) implique l'absence de hauteur. Ainsi, les sons bruités appartenant à la classe des sons bruités pseudo-périodiques sont caractérisés par une forte valeur de l'ACR, légèrement inférieur à 1.

Nombre de sinusoïdes

Nous avons déjà largement présenté ce descripteur dans la section 8.2 du chapitre 8 précédent.

Algorithme de classification général

Nous proposons donc un ensemble de trois descripteurs audio pour notre modèle de classification audio. Ces trois descripteurs sont : le rapport d'autocorrélation, le nombre de sinusoïdes et la fréquence des impulsions. Un marqueur est associé à chaque descripteur, ce marqueur peut prendre uniquement deux valeurs : 0 ou 1. La méthode de classification théorique, développée dans la section suivante, permet d'affecter une valeur à chaque marqueur. Cette valeur est égale à 1 si le descripteur estimé indique que le son analysé appartient au groupe qui correspond à ce descripteur. En revanche, la valeur 0 est affectée au marqueur.

À partir des valeurs prises par ces marqueurs associés à chaque descripteur, les règles suivantes sont appliquées pour classifier un son dans une des classes que nous avons proposées :

1. Si le marqueur du nombre de sinusoïdes est égal à 1, le marqueur relatif à la fréquence des impulsions est testé. S'il est égal à 1, alors le son analysé est inclus dans la classe C_{is}. S'il est égal à 0, alors le son analysé est inclus dans la classe C_s.
2. Si le marqueur du nombre de sinusoïdes est égal à 0, le marqueur de l'ACR est testé. S'il vaut 1, alors le son analysé est inclus dans la classe C_{ip}.
3. Si le marqueur de l'ACR vaut 0, le marqueur de la fréquence des impulsions est testé. S'il est égal à 1, alors le son considéré est inclus dans la classe C_i.
4. Si le marqueur de la fréquence des impulsions est égal à 0, alors le son considéré est inclus dans la classe C_c.

À la fin de l'algorithme, chaque son bruité est affecté à une classe. La condition est de définir de manière statistique la valeur du marqueur de chaque descripteur. Il s'agit d'un point essentiel de la méthode générale, qui est présenté dans la section suivante.

9.3 Méthode de classification

Après avoir présenté les classes de sons bruités ainsi que leurs descripteurs associés, nous allons décrire, dans cette partie, la méthode de classification statistique supervisée des sons bruités basée sur le modèle de décision bayésien. Chaque groupe, défini dans la section 9.2, est entraîné afin d'estimer les paramètres statistiques pour chacun d'entre eux. Plutôt que de mettre en place un schéma de classification à six classes, nous proposons une méthode, que nous pouvons qualifier de *une classe contre le reste*, qui classifie indépendamment chaque fenêtre d'analyse audio selon impulsif / non-impulsif, pseudo-périodique / non-pseudo-périodique et sinus / non-sinus. Ceci nous permet d'exploiter le schéma de classification à deux classes que nous avons déjà utilisé pour la détection des frontières de scènes (voir chapitre 6).Tous les descripteurs audio pour chaque fenêtre d'analyse sont calculés. Puis chaque vecteur de descripteurs associé à chaque fenêtre d'analyse est injecté dans le modèle de décision statistique que nous allons présenter.

Dans ce problème, nous associons une variable stochastique, x, au descripteur audio qui correspond au groupe considéré. Nous considérons aussi deux hypothèses, H_1 et H_2, telles que :

- H_1 : La fenêtre d'analyse courante appartient au Groupe considéré et
- H_2 : La fenêtre d'analyse courante n'appartient pas au Groupe considéré

H_1 et H_2 réalisent une partition de l'ensemble des hypothèses ($Pr(H_1) + Pr(H_2) = 1$).

Répétant les déductions du chapitre 6, et d'après les équations (109) à (114), nous obtenons le rapport de vraisemblance suivant pour chaque Groupe considéré :

$$\frac{L_1}{L_2} = \frac{P(x/H_1)}{P(x/H_2)} \tag{143}$$

Afin de pouvoir déterminer la fonction de densité de probabilité, $p(x/H_k)$, la mieux adaptée, une première phase d'apprentissage est nécessaire. Nous modélisons les fonctions de densité de probabilité comme des lois (distributions) statistiques qui correspondent le mieux avec la forme des histogrammes des valeurs obtenus lors de cette phase d'apprentissage. L'étape suivante consiste à suivre le développement de la méthode de décision statistique basée sur l'estimation du rapport de vraisemblance (voir sections 8.3 et 6.1), à savoir considérer le logarithme de ce rapport de vraisemblance (voir équation (177)). Ainsi, la décision finale est prise suivant la règle suivante :

$$M = log(\frac{L_1}{L_2}) \begin{array}{l} > \\ < \end{array} log(1) \begin{array}{l} \Rightarrow H_1 \\ \Rightarrow H_2 \end{array} \tag{144}$$

où "$=> H_k$" signifie que l'hypothèse H_k est la plus probable.

Par conséquent, notre problème consiste maintenant à appliquer la règle de décision (équation (178)) pour chacun des groupes G_p, G_i et G_s que nous avons définis dans la section 9.2. Ainsi, nous pouvons décider de manière indépendante l'appartenance, ou non, à chacun des groupes comme nous l'avons expliqué dans la partie 9.2.

9.4 Apprentissage et distributions utilisées

Comme nous l'avons précisé dans la partie précédente, notre schéma de classification statistique nécessite une phase d'entraînement afin d'estimer les distributions statistiques les mieux adaptées à la distribution des descripteurs, ainsi que les paramètres de ces distributions. La difficulté qui se pose est de choisir un corpus audio le plus représentatif possible de l'ensemble des sons bruités naturels. En effet, les avis divergent entre individus humains en ce qui concerne la détermination du type (de la classe) d'un son bruité naturel. C'est la raison pour laquelle nous avons décidé de constituer un corpus audio d'apprentissage avec des sons bruités *synthétiques* pour lesquels nous contrôlons leurs propriétés perceptives.
Dans la première sous-partie, nous présentons rapidement le modèle de synthèse de ces sons. Puis, nous introduisons les fonctions de densité de probabilité, que nous avons estimées, associées aux descripteurs audio. Enfin, dans la dernière sous-partie, nous montrons le bien fondé du choix de ces fonctions de densité de probabilité.

Méthode de synthèse des sons bruités

Le modèle classique de synthèse des sons bruités ne permet pas à l'utilisateur de contrôler les paramètres perceptifs si ce n'est la couleur, qui est fortement corrélée à l'enveloppe spectrale. Le modèle CNSS (Colored Noise by Sum of Sinusoids) a été introduit dans le but de pouvoir gérer les autres paramètres tels que l'harmonicité, l'impulsivité, la densité spectrale, etc... [Han03, HDC05]. De plus, du fait que ces paramètres dépendent de paramètres statistiques, un grand nombre de différents sons ayant les mêmes propriétés perceptuelles peuvent être synthétisés.

Le modèle CNSS [Han03] permet de représenter les sons bruités par une somme de sinusoïdes de courte durée, dont l'amplitude est fixée en fonction de l'enveloppe spectrale. Les fréquences et les phases sont des variables aléatoires. Seuls quelques paramètres du modèle permettent à l'utilisateur de contrôler les propriétés perceptives suivantes :

- L'impulsivité : un paramètre du modèle définit l'intervalle dans lequel sont déterminées de manière aléatoire les phases des sinusoïdes. Il est possible de produire un son bruité qui ne contient pas d'impulsions (bruit blanc) mais aussi un son bruité composé d'impulsions périodiques (ou non) avec différentes amplitudes.
- La périodicité : un paramètre du modèle contrôle la différence entre deux fréquences successives. Si cette différence (ou sa moyenne) correspond à une fréquence audible, alors une hauteur est perçue. Sinon aucune hauteur n'est perçue.

Le CNSS modèle a été développé dans [HBDC02]. Il nous permet de synthétiser 900 sons bruités d'une durée de 1 seconde et échantillonnés à 44100 Hz. Chaque son a une propriété particulière et peut être classifié dans un des groupes définis (G_p, G_i et G_s). Nous avons, ainsi, généré un corpus audio d'une durée totale de 15 minutes pour chacun des groupes et non-groupes.

Fonctions de densité de probabilité

L'étape d'apprentissage induit la définition des lois de distribution statistique pour chaque groupe et non-groupe. Dans beaucoup de travaux de classifications statistiques les auteurs

considèrent les distributions des descripteurs comme suivant la loi normale [TC04]. Cependant, des expérimentations ont montré que ces lois n'étaient pas adaptées pour tous les groupes. C'est pourquoi, nous avons décidé d'opter pour une loi de distribution mieux adaptée aux caractéristiques de chaque descripteur :

- Si le descripteur audio est défini de telle sorte qu'il admette une borne inférieure, et si cette borne correspond à une valeur maximale pour une fonction de densité de probabilité alors la loi exponentielle est une bonne candidate :

$$p(x) = \lambda \exp -(\lambda x) \text{ avec } x > 0 \tag{145}$$

  Si le domaine de définition du descripteur audio admet une borne supérieure, alors nous considérons le symétrique de la loi exponentielle.
- Si le descripteur audio est défini de telle sorte qu'il admette une borne inférieure, et si cette borne ne correspond pas à une valeur maximale pour une fonction de densité de probabilité alors la loi de Rayleigh est choisie :

$$p(x) = \frac{x}{\sigma^2} \exp -(\frac{x^2}{2\sigma^2}) \text{ avec } x \geq 0 \tag{146}$$

  avec σ l'écart-type.
  Si le domaine de définition du descripteur audio admet une borne supérieure, alors nous considérons le symétrique de la loi de Rayleigh.
- Si le descripteur audio est défini de telle sorte qu'il n'admette pas de borne supérieure (respectivement inférieure), la loi normale est considérée :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[\frac{-(x-\mu)^2}{(2\sigma^2)}] \tag{147}$$

  avec μ la moyenne des échantillons de x et σ l'écart-type.

Par conséquent, une loi de distribution est associée à chaque groupe et non-groupe. Le tableau de la figure 47 résume les lois de distributions associées à chaque groupe (respectivement non-groupe).

Groupe	**Loi statistique**
Sinusoïdes	Gauss
Non-sinusoïdes	Rayleigh
Pseudo-périodique	Rayleigh
Non-pseudo-périodique	Rayleigh
Impulsif	Gauss
Non-impulsif	Exponentielle

FIG. 47 – Groupes (resp. non-groupes) des sons bruités et leur fonction de densité de probabilité associée.

Résultats expérimentaux pour l'entraînement

Au cours de la phase d'apprentissage nous avons estimé les fonctions de densité de probabilité à associer à chaque descripteur audio proposé. Dans une première sous-partie, nous allons rappeler le déroulement de la phase d'apprentissage. Puis les sous-parties suivantes présenteront la justification des lois de distribution choisies pour chacun des groupes.

Méthode d'apprentissage

Les descripteurs audio que nous avons proposés sont extraits pour chacune des fenêtres d'analyse des sons bruités synthétiques de la base d'apprentissage. Pour chaque groupe et non-groupe, la moyenne et la variance de ces descripteurs sont calculées et ces deux valeurs induisent la définition des six fonctions de densité de probabilité associées à chaque groupe et non-groupe (voir tableau figure 47). Dans le but de minimiser les erreurs de classification et ainsi de valider nos estimations concernant le choix des fonctions de densité de probabilité, nous cherchons à maximiser la probabilité de classification correcte. Naturellement, la probabilité de classification correcte et la probabilité d'erreur de classification sont complémentaires. La probabilité de classification correcte peut-être définie par l'équation suivante :

$$P(\text{correct}/x) = \sum_{i=1}^{S} \int_{R_i} p(x/H_i)P(H_i)dx \tag{148}$$

avec S le nombre de classes considérées, R_i la i-ème région définie plus bas, $p(x/H_i)$ une fonction de densité de probabilité et $P(H_i)$ la probabilité de la classe H_i relative à la région R_i.

Dans notre problème, nous considérons deux classes, $S = 2$, et $P(H_1) = P(H_2) = \frac{1}{2}$. Ainsi, l'équation (148) précédente devient :

$$P(correct/x) = \frac{1}{2} \left[\begin{array}{c} \int_{R_1} p(x/H_1)dx \\ + \int_{R_2} p(x/H_2)dx \end{array} \right] \tag{149}$$

La figure 48 montre la manière de définir les régions R_i.

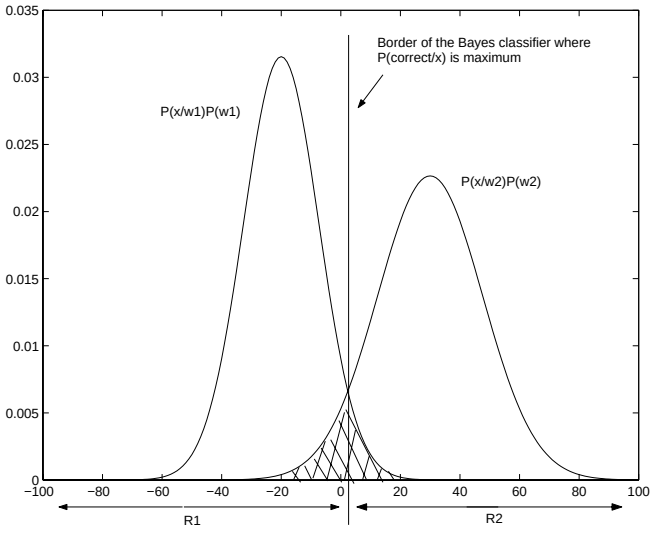


Fig. 48 – Illustration de la définition des régions R_i

Dans la sous-partie suivante, nous présentons les résultats obtenus pour l'estimation de la probabilité de bonne classification pour chaque paire Groupe / Non-Groupe.

Résultats pour Impulsif / Non-Impulsif

La probabilité de bonne classification expérimentale que nous avons obtenue est égale à 0.99. La figure 49 représente la fonction de densité estimée pour ce groupe (respectivement non-groupe) et illustre le résultat obtenu. La courbe de la distribution expérimentale associée au groupe non-impulsif a une forme très fine et pointue. Cette caractéristique peut être expliquée par le faible nombre de valeurs récoltées pour la fréquence des impulsions des sons bruités qui ne sont pas considérés comme impulsifs. En effet, la valeur de la fréquence des impulsions est théoriquement nulle dans ce cas. Cependant, certaines valeurs très faibles, au voisinage de zéro, peuvent élargir la forme de la courbe de la loi de distribution considérée.

La courbe des valeurs de la distribution pour les sons bruités impulsifs a une forme plus élargie que la courbe des valeurs de la distribution pour les sons bruités non-impulsifs. La fréquence des impulsions peut prendre différentes valeurs dans l'intervalle des fréquences audibles. Nous pensons, donc, que la loi de distribution théorique peut-être uniforme. De plus, il est possible d'estimer une distribution de valeurs uniformes par une loi normale.

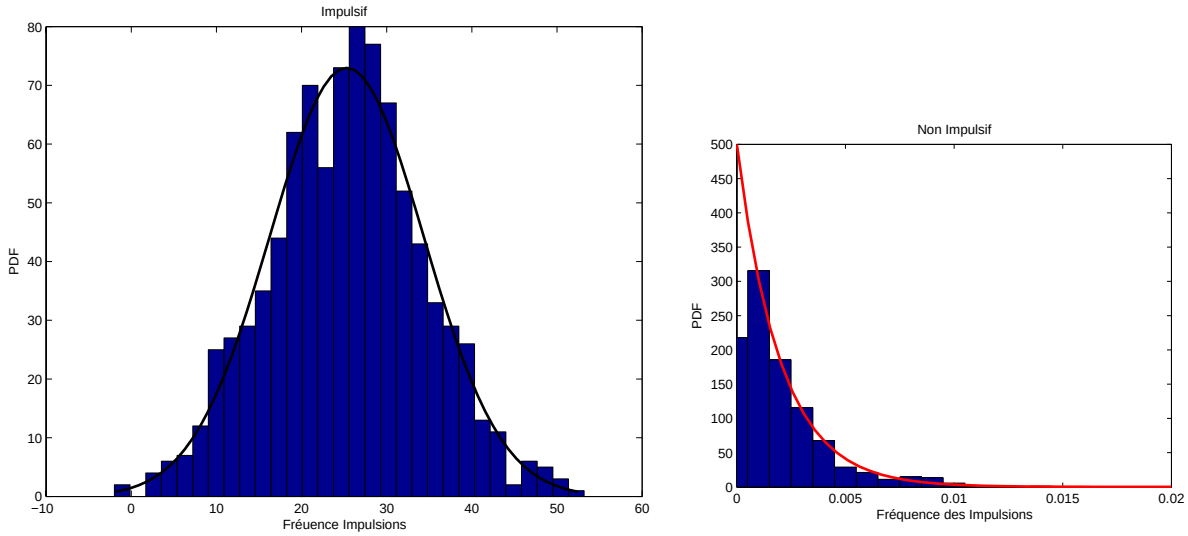


FIG. 49 – Fonctions de densité de probabilité expérimentales pour les groupes Impulsif et Non-impulsifs.

Résultats pour Avec-Sinusoïdes / Non-Avec-Sinusoïdes

La probabilité expérimentale de bonne classification obtenue est égale à 0.96. La figure 50 montre la fonction de densité de probabilité estimée pour ce groupe (respectivement non-groupe) et illustre le résultat que nous avons obtenu. La distribution pour le groupe non-avec-sinusoïdes semble similaire à celle de groupe non-impulsif. Les explications sont, donc, les mêmes : les sons bruités non-avec-sinusoïdes sont caractérisés par un très faible nombre de sinusoïdes. Si la méthode d'analyse était parfaite alors ce nombre serait nul. Dans notre cas, de légères imprécisions sont dues à certaines incertitudes de la méthode d'analyse proposée. Cela se traduit par un élargissement de la forme de la courbe des valeurs de la distribution.

La distribution pour le groupe avec-sinusoïdes a les mêmes caractéristiques que celle du groupe impulsif. De la même manière, nous supposons ici que la distribution est uniforme.

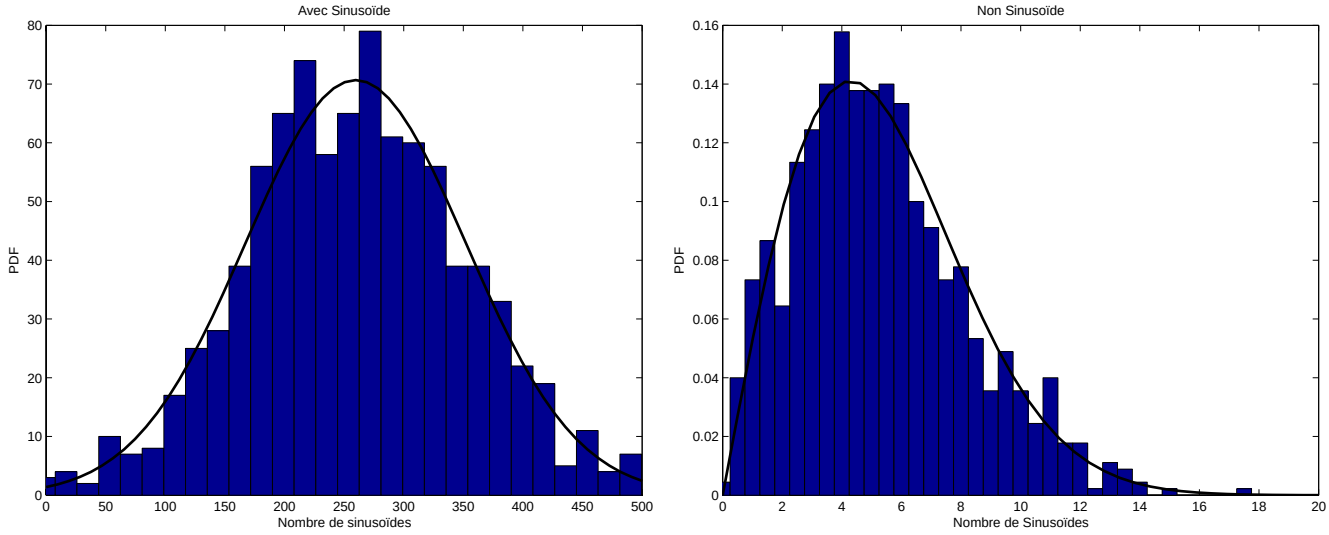


Fig. 50 – Fonctions de densité de probabilité expérimentales pour les groupes Avec-Sinusoïdes et Non-Avec-Sinusoïdes.

Résultats pour Périodique / Non-Périodique

La probabilité expérimentale de bonne classification obtenue est égale à 0.99. La figure 51 montre la fonction de densité de probabilité estimée pour ce groupe (respectivement non-groupe) et illustre le résultat que nous avons obtenu. La distribution pour les groupes périodique et non-périodique sont approximativement symétriques. Cette caractéristique s'explique par les valeurs du rapport d'autocorrélation qui sont comprises entre 0 et 1. Les sons bruités non-périodiques sont caractérisés par des valeurs expérimentales du rapport d'autocorrélation comprises dans l'intervalle $[0; 0.3]$ tandis que les sons périodiques sont caractérisés par des valeurs expérimentales du rapport d'autocorrélation comprises dans l'intervalle $[0.7; 1]$.

9.5 Résultats et expérimentations

Diverses expérimentations ont été réalisées afin d'une part de valider l'ensemble des descripteurs proposés et d'autre part, de mettre en avant les poins forts et les points faibles de la méthode de classification statistique et supervisée des sons bruités naturels. Nous avons aussi mis en avant la pertinence du choix des classes de sons bruités.

L'ensemble de ces tests et des résultats obtenus pour ces travaux sont exposés dans le chapitre 12 de la partie III.

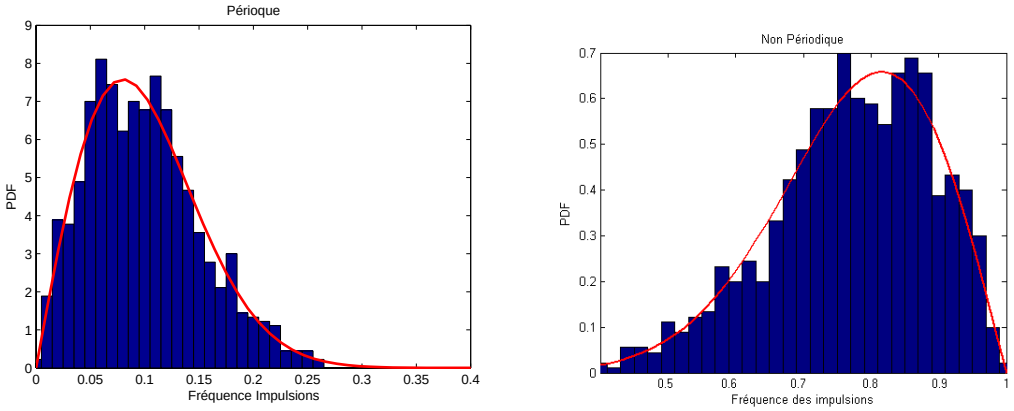


FIG. 51 – Fonctions de densité de probabilité expérimentales pour les groupes périodiques et non-périodiques.

Quatrième partie

Expérimentations et résultats

Chapitre 10

Indexation cross-média

Dans ce chapitre, nous présentons la phase d'expérimentations que nous avons menée dans le cadre de nos travaux sur l'indexation cross-média.

Dans ce travail expérimental, nous avons mis au point un batterie de tests afin de valider le modèle de scène que nous avons proposé. Pour cela, nous devons nous affranchir des erreurs provenant des détecteurs des flux audio et vidéo en considérant les vérités terrain de ces approches. En fin de compte, notre but est d'une part de vérifier le bien-fondé de notre modèle, en le confrontant à un corpus audiovisuel de test afin de vérifier si notre hypothèse de départ est valide ou non. D'autre part, nous devons évaluer les performances de la méthode de décision statistique que nous avons développée.

Pour l'ensemble de ces tests, nous avons utilisé une base de test vidéo au format MPEG-2 d'une durée totale de 190 minutes qui nous a été fournie par Philips Research et composée de la manière suivante :

- 60 minutes de séries réparties en deux programmes différents : une série extraite de la chaîne RTL4 (Hollandaise) et une autre de la chaîne RTL (Allemande).
- 30 minutes de documentaire provenant de la chaîne BBC2 (Britannique).
- 100 minutes de film issues de la chaîne française TV5.

Le tableau suivant 52 donne un certain nombre de détails concernant le corpus audiovisuel considéré comme le nombre de frontières de scène ainsi que le nombre de plans de montage réels des échantillons qui composent notre base de test. Conformément à notre modèle, nous avons

Genre	Durée (mins)	Nbre de plans de montage	Nbre de silences	Nbre de scènes
Séries	60	921	177	50
Documentaire	30	410	58	13
Film	100	1174	225	27
Total	190	2505	460	90

FIG. 52 – Vérité terrain des échantillons du corpus de test

annoté manuellement toutes les transitions vidéo, tous les silences et toutes les frontières de scène qui correspondent à notre modèle. Il nous semble judicieux de donner des commentaires sur l'annotation de la vérité terrain. Pour les transitions vidéo, un parcours image par image du flux nous a permis de localiser les cuts. Le numéro de l'image que nous avons reporté

dans la vérité terrain correspondent à la dernière image du plan considéré. Concernant les transitions audio (silences), nous avons visualisé l'amplitude du signal audio en fonction du temps afin de localiser les zones pour lesquelles le niveau sonore est inférieur au seuil d'audition (voir diagramme de Fletcher 12). Pour ce qui est de l'annotation des scènes, nous avons déjà mentionné la difficulté de l'annotation sémantique des scènes dans le chapitre 5. Étant donné que nous avons travaillé sur des contenus artistiques et tout en étant conscient des limites d'une telle annotation, nous avons choisi de faire une annotation sémantique basée uniquement sur le critère de la cohérence du lieu (voir section 5.3.1). Nous avons ainsi annoté 137 frontières de scène. Dans un premier temps, parmi les frontières annotées nous avons sélectionné uniquement celles qui correspondaient à notre modèle (vidéo cut + silence audio). Le nombre total de scènes correspondant à ce modèle est de 90. Les résultats de ces annotations sont résumés dans le tableau 52.

Les détecteurs de silence et des frontières de plan de montage ainsi que le détecteur de scène ont été testés à partir de la même base de données.

Enfin, ce chapitre est organisé de la manière suivante. La première partie est réservée à la présentation des performances de l'analyse des flux audio et vidéo. Dans une deuxième partie, nous présentons les différents tests que nous avons menés. Nous exposons les résultats des performances du détecteur de frontières de scène sur l'ensemble du corpus mélangé dans la troisième partie. Puis ces mêmes résultats ventilés par genre vidéo, présents dans notre corpus vidéo de test, sont présentés dans la dernière partie de ce chapitre.

Les résultats présentés dans ce chapitre ont été obtenus suite à une collaboration scientifique avec M. Jan Nesvadba (Philips Research NatLab). Ce qui implique que les résultats obtenus sont exploitables par les deux parties.

10.1 Analyse des flux audio et vidéo

Dans cette partie, nous donnons rapidement les performances obtenues à partir des détecteurs de silence et de frontières de plans de montage. Nous présentons ces résultats de manière relativement sommaire car nous nous focalisons, dans ces travaux, sur la méthode de fusion de l'information multimédia. Comme mesure de performances de ces détecteurs nous avons retenu les indicateurs classiques de rappel et de précison dont les formules respectives sont rappelées ci-dessous :

$$Rappel = \frac{N_{correctementDétecté(e)s}}{N_{véritéTerrain}} \quad (150)$$

$$Prcision = \frac{N_{correctementDétecté(e)s}}{N_{correctementDétecté(e)s} + N_{faussementDétecté(e)s}} \quad (151)$$

avec N_x le nombre total de x.

Les performances du détecteur de silences sont passées en revue dans la première sous-partie. Pour ce qui est des performances de l'analyse vidéo, elles sont résumées dans la deuxième sous-partie.

Détection des silences

La détection des silences dans les domaines compressé et non compressé a été testée sur le corpus de test présenté ci-dessus. Compte tenu de la simplicité de la règle de décision, il nous a été difficile de régler le seuil (voir équation 125).

Par rapport à la méthode de calcul du seuil (cf chap 7), nous l'avons ajusté de façon empirique. Ces seuils ont été déterminés de manière expérimentale pour chaque document . Cela s'explique par le fait que les enregistrements de ces contenus ne produisent pas un volume audio normalisé ce qui signifie que des variations d'intensité sont perceptibles d'un contenu à l'autre.

Les tableaux des figures 53 et 54 résument les performances obtenues dans les domaines compressé et non compressé respectivement :

Domaine compressé :

Genre Vidéo	Nbre silences attendus	Rappel	Précision
Série	177	96.70%	96.70%
Documentaire	58	92.20%	86.60%
Film	225	93.70%	92.80%

Fig. 53 – Résultats de la détection des silences (domaine compressé)

Domaine non compressé :

Genre Vidéo	Nbre silences attendus	Rappel	Précision
Série	177	82.00%	55.00%
Documentaire	58	73.00%	54.00%
Film	225	78.00%	54.00%

Fig. 54 – Résultats de la détection des silences (domaine non compressé)

Les meilleures performances sont observées dans le domaine compressé du fait de la qualité du signal qui a été filtré lors de la phase d'encodage et par la simplicité de la règle de décision ainsi que la difficulté à fixer un seuil. Quelques erreurs de détections sont observées dans le domaine non compressé à cause du manque de précision temporelle du détecteur. De plus, après décompression, l'ensemble des fréquences ayant été quantifiées lors du codage nous perdons de l'information notamment sur les faibles valeurs d'amplitude qui sont quantifiées à 0 générant ainsi un grand nombre de fausses alarmes. La nécessité d'utiliser un seuil assez faible dans le domaine non compressé pour la détection des silences bruités est à l'origine de la relative faible précision observée du détecteur.

Détection des frontières de plans de montage

De la même manière que dans les cas des silences, la méthode de caractérisation des frontières de plans de montage, propriétaire du projet CASSANDRA [Cas] a été mise à l'épreuve sur la même base de test (définie ci-dessus). Le tableau de la figure 55 énumère les performances obtenues.

Genre Vidéo	Vérité terrain	Rappel	Précision
Série	921	98.20%	98.30%
Documentaire	410	98.30%	98.70%
Film	1174	96.60%	96.60%

FIG. 55 – Performances du détecteur de frontières de plans de montage

Les plus petites valeurs de rappel et précision sont obtenues sur les contenus de type film. Pour ce qui est du rappel, son score s'explique par la présence de transitions graduelles de courte durée que nous avons assimilées à des transitions de type coupure lors de l'annotation manuelle du contenu. Dans le cas de la précision, ceci s'explique par le fait que le détecteur proposé est, de part son principe, basé sur l'estimation du mouvement obtenue lors de la phase d'encodage du contenu, sensible aux mouvements de forte amplitude.

10.2 Méthodes de détection des frontières de scène

Cette partie est consacrée à la description des schémas de tests prévus pour valider les performances de la méthode de décision statique mise en œuvre pour la détection des frontières de scènes. Ces mêmes scènes sont, selon notre hypothèse, caractérisées par deux frontières successives. Ces frontières sont, toujours selon notre hypothèse de départ, matérialisées par la présence conjointe d'un silence et d'une frontière de plan de montage de type coupure dans le même instant de temps voire avec léger décalage temporel (jitter). Dans notre schéma de décision statistique, nous nous permettons une exception : dans le cas d'une variance nulle nous fixons la valeur de la variance à 1 par défaut. Cette règle n'est pas très préjudiciable dans la prise de décision car l'écart entre ces valeurs est négligeable dans le contexte de nos travaux.

Une description complète de ces méthodes de tests est donnée dans les deux sous-parties qui suivent. D'une part, une règle de décision simple consistant à utiliser un seuil fixe pour évaluer les valeurs de jitter. La valeur de ce seuil est exprimée en terme de nombre d'images. Une image correspond à 1/25 de seconde dans notre cas, 25 images par seconde dans les vidéos de notre base de test. D'autre part, une règle de décision plus robuste qui utilise la méthode de décision statique, développée dans le chapitre 6 est aussi utilisée. Les performances obtenues avec ces deux méthodes sont comparées.

Règle de décision simple par seuillage

Nous avons décidé d'utiliser une règle de décision simple tout d'abord de part son faible coût en terme de ressource. Une valeur de seuil est fixée au départ. Ce qui signifie que toutes les frontières de plan de montage, associées aux valeurs de jitter inférieures ou égales en valeur absolue au seuil, seront considérées comme des frontières de scène. Évidemment, si une valeur de jitter est strictement supérieure au seuil alors nous la rejetons et nous passons à la suivante.

Règle de décision robuste par approche statistique

Dans cette sous-partie, nous expliquons les détails pratiques des expériences menées en utilisant la règle de décision robuste basée sur l'utilisation de la méthode de décision bayésienne. Ce schéma de décision se décompose suivant deux principales phases :

- **la phase d'apprentissage** : cette phase consiste à apprendre les paramètres statistiques relatifs aux classes "Changement de scène" et "Non-changement de scène" par rapport à une base d'apprentissage manuellement annotée et
- **la phase de test** : dans cette phase, deux séries d'expérimentations ont été effectuées. La première étant la caractérisation de la qualité de l'apprentissage et le seconde la caractérisation de la qualité de détecteur.

L'entraînement du système a été réalisé de la manière suivante. Nous avons au préalable annoté l'ensemble de notre corpus vidéo, soit 190 minutes de contenus vidéo. Nous avons décidé de partager notre base de test en deux : la première moitié de chaque vidéo est destinée à entraîner le système alors que l'autre moitié est utilisé comme base de test. Pour chaque transition vidéo manuelle annotée comme étant une frontière, la valeur du jitter associée est récupérée et ajoutée dans un ensemble de données associée à la classe "Changement de scène". La même procédure est appliquée pour les valeurs de jitter qui correspondent à la classe "Non-changement de scène".

Après quoi, pour chaque ensemble de données, $X_j = \{x_{j1}, \ldots, x_{jN_j}\}$ $(j = 1, 2)$: "changement de scène" et "non-changement de scène"), la moyenne, μ_j, et la variance, σ_j^2, des valeurs sont calculées.

Comme expliqué ci-dessus, la seconde moitié des vidéos sont considérées comme un ensemble de test non labellisé. Cette autre moitié est utilisée pour mettre en évidence les performances de notre méthode de décision statistique basée sur les paramètres gaussiens (moyenne et variance) estimés lors de la phase d'apprentissage. Nous obtenons donc deux formes de résultats pour les performances de notre méthode de décision statistique :

- les performances obtenues sur la base d'apprentissage (première partie de la vidéo) et
- les performances obtenues sur la base de test (seconde moitié de la vidéo).

Enfin, pour être en adéquation avec notre hypothèse de départ, seules les frontières de scènes qui correspondent à notre modèle, transition vidéo de type coupure associée à un silence, sont prises en considération pour le calcul des valeurs de rappel et de précision.

10.3 Résultats globaux sur l'ensemble du corpus de test

Classe "Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

- Variance $(\sigma_1^2) = 0{,}76$
- Moyenne $(\mu_1) = 0{,}11$
- Probabilité de classe = *Nombre total de frontières de scènes / Nombre total de frontières de plans de montage* $= 90/2505 = 0,03$

Classe "Non-Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

- Variance (σ_2^2) = 321102,70
- Moyenne (μ_2) = 26,67
- Probabilité de classe = *Nombre total de non-frontières de scènes / Nombre total de frontières de plans de montage* = 2415/2505 = 0,97.

Deux méthodes sont comparées : la règle de décision simple utilisant un seuil fixe et la méthode de décision statistique.

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
47	18	29	128	38,29%	12,33%

FIG. 56 – Résultats de la détection simple, seuil $+/-3$, sur l'ensemble du corpus de test

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
47	20	27	194	42,55%	9,35%

FIG. 57 – Résultats de la détection simple, seuil $+/-10$, sur l'ensemble du corpus de test

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
47	23	24	246	48,94%	8,55%

FIG. 58 – Résultats de la détection simple, seuil $+/-21$, sur l'ensemble du corpus de test

Base	Vérité terrain	Correctes	Manquées	Fausses alarmes	Rappel	Précision
Apprentissage	43	35	8	95	81,40%	26,93%
Test	47	15	32	127	31,92%	10,57%

FIG. 59 – Résultats de la détection statistique sur l'ensemble du corpus de test

Les tableaux des figures 56, 57, 58 et 59 montrent les résultats, respectivement pour la règle de décision simple avec un seuil de décision égal à 3, 10 et 21 images vidéos et pour la règle de décision statistique bayésienne appliquée à l'ensemble du corpus vidéo, tous genres confondus.

Néanmoins, ces résultats ne reflètent pas les réelles performances de notre détecteur des frontières de scènes. En effet, les performances sont perturbées par les erreurs provenant à la fois de l'analyse des flux audio et vidéo. C'est pourquoi, dans les tableaux des figures ci-dessous nous donnons les performances pures pour la détection des frontières de scène. Pour cela, nous nous affranchissons des perturbations liées aux erreurs dans l'analyse des flux audio et vidéo en considérant la vérité terrain et les mesures de rappel et précisions ajustées pour ces détecteurs. Dans un premier temps, nous introduisons la vérité terrain ajustée ($VT_{ajustée}$). Cette dernière comporte toutes les frontières des scènes présentes dans la vérité terrain initiale sauf celles pour lesquelles le détecteur vidéo a manqué le changement de plan et/ou le détecteur audio a manqué un silence. La valeur de rappel ajustée est calculée par rapport à la vérité terrain

ajustée :

$$Rappel_{ajusté} = \frac{N_{correctementDetectées}}{VT_{ajustée}} \tag{152}$$

avec N le nombre de frontières de scènes.
Par rapport aux détecteurs réels, nous introduisons la "précision ajustée". Nous considérons le nombre total de frontières de scène faussement détectées auquel nous retranchons le nombre total de frontières de scène faussement détectées à cause d'une fausse alarme provenant des détecteurs des flux audio et vidéo. Cette dernière valeur est dénotée par $N_{faussementDéetectéesAjustée}$. Ainsi,

$$Précision_{ajustée} = \frac{N_{correctementDetectées}}{N_{correctementDetectées} + N_{faussementDéetectéesAjustée}} \tag{153}$$

Ces valeurs ajustées correspondent en fait à la détection idéale sur un sous-ensemble des frontières des scènes. Ces mesures sont présentées dans les tableaux ci-dessous.

Dans la suite de ce document les abbréviations R_A, R_B, R_C, R_D, R_E, R_F, R_G, R_H et R_I correspondent à :

- R_A : Vérité terrain,
- R_B : Nombre total de scènes manquées,
- R_C : Nombre de silences audio manqués relatifs à une frontière de scène ,
- R_D : Nombre de transitions vidéo manquées liées à une frontière de scène,
- R_E : Nombre total d'erreurs provenant des détecteurs audio et vidéo (E = C + D),
- R_F : Vérité terrain ajustée (F = A - E),
- R_G : Nombre ajusté de frontières de scènes manquées,
- R_H : Nombre de scènes correctement détectées et
- R_I : Valeur de rappel ajustée (I = H / F).

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
47	29	20	0	20	27	9	18	66,66%

Fig. 60 – Valeur de rappel ajustée sur l'ensemble du corpus de test avec seuil fixé à $+/-3$

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
47	27	17	0	17	30	10	20	66,66%

Fig. 61 – Valeur de rappel ajustée sur l'ensemble du corpus de test avec seuil fixé à $+/-10$

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
47	24	14	0	14	33	10	23	69,69%

Fig. 62 – Valeur de rappel ajustée sur l'ensemble du corpus de test avec seuil fixé à $+/-21$

Ces valeurs de rappel correspondent au mélange des classes "changement de scène" et "non-changement de scène". De la même manière, les tableaux des figures suivantes exhibent les valeurs de la précision pure du détecteur de frontières de scènes.

Pour la même raison que pour les tableaux précédents, voici la correspondance des lettres pour la suite de ce document :

Base	R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
Apprentissage	43	8	5	0	5	38	3	35	92,11%
Test	47	32	29	0	27	20	5	15	75,00%

FIG. 63 – Valeur de rappel ajustée sur l'ensemble du corpus avec la méthode de décision bayésienne

- P_A : Vérité terrain,
- P_B : Nombre de frontières de scène faussement détectées,
- P_C : Nombre de silences faussement détectés,
- P_D : Nombre de transitions vidéo faussement détectées,
- P_E : Nombre total de fausses détections des frontières de scène provenant des détecteurs audio et vidéo,
- P_F : Nombre ajusté de frontières de scènes faussement détectées,
- P_G : Nombre de frontières de scène correctement détectées et
- P_H : Valeur de précision ajustée (H = G / (G + F)).

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
47	128	155	23	110	18	18	50,00%

FIG. 64 – Valeur de précision ajustée sur l'ensemble du corpus de test avec seuil fixé à $+/-3$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
47	194	155	23	165	29	20	40,81%

FIG. 65 – Valeur de précision ajustée sur l'ensemble du corpus de test avec seuil fixé à $+/-10$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
47	246	155	23	170	76	23	23,23%

FIG. 66 – Valeur de précision ajustée sur l'ensemble du corpus de test avec seuil fixé à $+/-21$

Base	P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
Apprentissage	43	95	80	18	79	16	35	68,63%
Test	47	127	215	22	111	16	15	48,38%

FIG. 67 – Valeur de précision ajustée sur l'ensemble du corpus avec la méthode de décision bayésienne

Au regard des valeurs obtenues pour les performances pures de segmentation en scènes avec l'utilisation des détecteurs audio et vidéo dans leur version "idéale", nous notons une amélioration très nette des valeurs de rappel et de précision sauf dans le cas de la décision avec un seuil très faible. Ainsi, nous avons montré que les performances des détecteurs multimédia jouent un rôle important dans le bon déroulement du principe de segmentation en scènes et il est donc important de faire appel à des méthodes robustes. Si nous ne tenons pas compte de

tous les cas d'erreurs ou de fausses alarmes provenant des détecteurs audio et vidéo, alors nous pouvons justifier les valeurs de rappel et de précision par le mélange des classes. En fait, une forte valeur de jitter correspondant à un changement de scène signifie que l'utilisateur a annoté ce changement de scène lorsqu'une transition vidéo n'était pas accompagnée d'un silence dans le flux audio. Au contraire, une faible valeur de jitter dans la classe "non-changement de scène" correspond à la présence d'une scène de dialogues.

Revenons sur les tableaux des figures 56 à 59 qui montrent les performances de la méthode avec les détecteurs audio et vidéo réels. Nous pouvons remarquer que le fait d'utiliser un simple seuil, égal à +/-3, pour la classification permet d'obtenir des résultats assez proches de ceux obtenus avec la règle de décision bayésienne. Au contraire, les performances sont moins bonnes si nous utilisons une valeur de seuil trop élevée. Concernant la règle de décision bayésienne, $31,92\%$ des frontières de scènes ont été correctement détectées. Cela peut être expliqué par un jitter trop élevé pour être considéré comme appartenant à la classe "changement de scène". Cette particularité peut s'expliquer de deux manières :

- par une mauvaise détection de la part du détecteur audio ou vidéo ou bien
- par la présence de réelles grandes valeurs de jitter pour lesquelles il nous est impossible de détecter les frontières de scène associées.

Malheureusement, nous obtenons de mauvaises valeurs de précision, seulement $10,57\%$. Ceci s'explique aussi pour deux raisons :

- une mauvaise détection de la part du détecteur audio ou vidéo ou bien
- un recouvrement trop important des distributions gaussiennes associées aux deux classes. En d'autres termes, cela signifie que dans certains cas de non-changement de scène nous sommes aussi en présence d'une valeur faible de jitter.

Ces faibles valeurs de précision sont aussi dues à la présence de scènes de dialogues. Scènes de dialogue dans lesquelles il est souvent possible d'observer l'alternance des prises de vues rapprochées des deux protagonistes caractérisée par la présence de transitions vidéo de type coupure et la présence de silence audio, d'où la présence abusive de frontières de scènes selon notre modèle. Après analyse des performances de la règle de décision simple et de décision bayésienne, nous pouvons voir que la règle de décision statistique donne des performances meilleures en terme de précision même si les valeurs de rappel sont légèrement inférieures. Ceci est dû, en partie, au fait que nous prenons en considération la probabilité des classes dans la règle de décision finale.

Le tableau de la figure 68 exprime les performances pures de notre détecteur de frontières de scènes en utilisant les valeurs manuelles de la segmentation audio et vidéo. Cette expérimentation est équivalente à celle qui consiste à ajuster les valeurs de la vérité terrain, de fausses alarmes et des erreurs de segmentation. Cependant, les résultats obtenus sont différents du fait que nous n'avons pas à ré-ajuster la vérité terrain, les 90 scènes sont prises en considération. Ici, nous remarquons que la méthode de la décision statistique permet d'obtenir une meilleure précision et est légèrement inférieure en rappel.

Les performances données par la table 68 sont meilleures que celles obtenues pour les valeurs ajustées de rappel et de précision données par les tables 60 à 67. Ceci peut être expliqué par le fait que dans les tables 60 à 67 le calcul des valeurs de rappel et de précision est basé sur la vérité terrain ajustée, tandis que les performances données par la table 68 sont

Règle	Vérité terrain	Correctes	Manquées	Fausses	Rappel	Précision
Jitter+/- 3	47	42	9	18	89,40%	70,00%
Jitter+/- 10	47	37	10	29	78,72%	56,06%
Jitter+/- 21	47	37	10	76	78,72%	32,74%
Statistique (Appr.)	43	40	3	16	93,02%	71,42%
Statistique (Test)	47	39	5	16	88,64%	70,90%

FIG. 68 – Performances idéales, sur l'ensemble du corpus, du classifieur à partir des valeurs manuelles de la segmentation audio et vidéo

obtenues avec la vérité terrain idéale.
La règle de décision simple rate 9 scènes associées à une faible valeur de jitter et 10 avec une grande valeur de jitter. La règle de décision statistique rate 5 scènes à cause du mélange des classes considérées. De plus, le système de décision statistique n'a que quelques valeurs de jitter comme base d'apprentissage, la première moitié des vidéos. De surcroît le début de chaque contenu audiovisuel est souvent composé de publicités, pour lesquelles notre modèle est très bien adapté d'où la présence de faibles valeurs de jitter pour l'entraînement du système. De ce fait, lors de la phase de test le système n'est pas capable de détecter les frontières de scènes avec un jitter un peu trop élevé. La précision est meilleure que lors des expérimentations précédentes mais elle reste assez faible tout de même, 70% dans le meilleur des cas pour la règle de décision simple et $70,90\%$ pour la règle de décision statistique. Ici aussi, la relative faible précision s'explique par la présence importante de scènes de dialogue qui, comme expliqué précédemment, génère un nombre important de fausses alarmes. C'est pourquoi, nous pensons qu'il serait utile d'inclure au système une méthode de reconnaissance des scènes de dialogue afin d'améliorer nettement la précision de notre méthode de détection.

Il nous paraît assez évident que notre modèle de scène est dépendant du type de contenu audiovisuel. Cependant, il est maintenant intéressant d'analyser ses performances pour chaque type de genre vidéo afin de dégager les types de contenus pour lequel notre modèle est mieux adapté, de même pour les cas où il est moins bien adapté.

10.4 Résultats par genre vidéo

Dans cette partie, nous exposons nos résultats ventilés par type de contenu vidéo afin de de dégager les types de contenus pour lequel notre modèle est mieux adapté, de même pour les cas où il est moins bien adapté. Les genres vidéo que nous avons à notre disposition sont les séries, les documentaires et les films. Ce sont d'ailleurs ces trois types de contenus qui vont constituer respectivement les trois sous-parties suivantes.

Série

Classe "Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

– Variance (σ_1^2) = 0,297

- Moyenne (μ_1) = -0,16
- Probabilité de classe = *Nombre total de frontières de scènes / Nombre total de frontières de plans de montage* = $50/921 = 0,05$

Classe "Non-Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

- Variance (σ_2^2) = 128126,56
- Moyenne (μ_2) = 13,93
- Probabilité de classe = *Nombre total de non-frontières de scènes / Nombre total de frontières de plans de montage* = $871/921 = 0,95$.

Deux méthodes sont comparées : la règle de décision simple utilisant un seuil fixe et la méthode de décision statistique.

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
26	13	13	75	50,00%	14,77%

Fig. 69 – Résultats de la détection simple, seuil $+/-3$, pour les contenus de type série du corpus de test

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
26	14	12	120	53,84%	10,41%

Fig. 70 – Résultats de la détection simple, seuil $+/-10$, pour les contenus de type série du corpus de test

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
26	16	10	150	61,54%	9,64%

Fig. 71 – Résultats de la détection simple, seuil $+/-21$, pour les contenus de type série de corpus de test

Base	Vérité terrain	Correctes	Manquées	Fausses alarmes	Rappel	Précision
Apprentissage	24	19	5	38	79,17%	33,34%
Test	26	10	16	81	38,46%	11,00%

Fig. 72 – Résultats de la détection statistique pour les contenus de type série du corpus

Les tableaux des figures 69 à 72 montrent les résultats respectifs de la détection simple avec seuil fixé à 3, 10 et 21 et de la détection statistique bayésienne pour les contenus de type série du corpus. De la même manière que dans la section précédente consacrée aux résultats des expériences menées sur l'ensemble du corpus tous genres confondus, nous allons maintenant considérer les performances de la détection des frontières de scènes avec les détecteurs audio et vidéo "idéaux". Nous allons aussi calculer les valeurs ajustées de rappel et de précision dans les tableaux des figures ci-dessous.

Au regard des résultats obtenus pour les valeurs de rappel et de précision pures avec les détecteurs audio et vidéo ajustés (tableaux 73 à 80), nous remarquons une amélioration

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
26	13	8	0	8	18	5	13	72,22%

FIG. 73 – Valeur de rappel ajustée sur les contenus de type série du corpus de test avec seuil fixé à $+/-3$

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
26	12	7	0	7	19	5	14	73,68%

FIG. 74 – Valeur de rappel ajustée sur les contenus de type série du corpus de test avec seuil fixé à $+/-10$

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
26	10	5	0	5	21	5	16	76,19%

FIG. 75 – Valeur de rappel ajustée sur les contenus de type série du corpus de test avec seuil fixé à $+/-21$

Base	R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
Apprentissage	24	5	5	0	5	19	0	19	100,00%
Test	26	16	13	0	13	13	3	10	76,92%

FIG. 76 – Valeur de rappel ajustée sur les contenus de type série du corpus avec la méthode de décision bayésienne

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
26	75	77	6	65	10	13	56,52%

FIG. 77 – Valeur de précision ajustée sur les contenus de type série du corpus de test avec seuil fixé à $+/-3$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
26	120	77	6	93	27	14	34,15%

FIG. 78 – Valeur de précision ajustée sur les contenus de type série du corpus de test avec seuil fixé à $+/-10$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
26	150	77	6	96	54	16	22,86%

FIG. 79 – Valeur de précision ajustée les contenus de type série du corpus de test avec seuil fixé à $+/-21$

sensible des valeurs de rappel et de précision. Les meilleurs performances, en termes de rappel et de précision, sont observées dans la cas de la décision statistique avec $93,34\%$ de précision et $76,92\%$ de rappel.

Revenons sur les tableaux 69 à 72, il est possible de remarquer que la règle de décision simple avec un seuil égal $= +/-21$ permet d'obtenir la meilleure valeur de rappel avec $61,54\%$. Tandis que la meilleure valeur de précision est obtenue par la méthode de décision

Base	P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
Apprentissage	24	38	53	4	37	1	19	95,00%
Test	26	81	162	8	79	2	10	83,34%

Fig. 80 – Valeur de précision ajustée sur les contenus de type série du corpus avec la méthode de décision bayésienne

simple $+/-3$ avec $14,77\%$.

Règle	Vérité terrain	Correctes	Manquées	Fausses	Rappel	Précision
Jitter+/- 3	26	21	5	10	80,77%	67,75%
Jitter+/- 10	26	21	5	27	80,77%	43,75%
Jitter+/- 21	26	21	5	54	80,77%	28,00%
Statistique (Appr.)	24	24	0	1	100,00%	96,00%
Statistique (Test)	26	23	3	2	88,46%	92,00%

Fig. 81 – Performances idéales, sur les contenus de type série du corpus, du classifieur à partir des valeurs manuelles de la segmentation audio et vidéo

Au regard des résultats fournis par la table 81 et obtenus à partir des valeurs manuelles des détecteurs audio et vidéo, nous pouvons remarquer une amélioration des valeurs de rappel et de précision par rapport aux performances obtenues pour les valeurs ajustées de rappel et de précision. Le cas optimal pour les valeurs de rappel et de précision est obtenu à l'aide de la méthode de décision statistique : $92,00\%$ pour la précision et $88,46\%$ pour le rappel. La décision simple garantit $88,77\%$ de rappel et $67,75\%$ de précision.

Documentaire

Classe "Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

- Variance (σ_1^2) = 0,00 (= 1,00 par défaut)
- Moyenne (μ_1) = 0,00
- Probabilité de classe = *Nombre total de frontières de scènes / Nombre total de frontières de plans de montage* = $13/410 = 0,03$

Classe "Non-Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

- Variance (σ_2^2) = 92251,60
- Moyenne (μ_2) = -17,45
- Probabilité de classe = *Nombre total de non-frontières de scènes / Nombre total de frontières de plans de montage* = $397/410 = 0,97$.

Deux méthodes sont comparées : la règle de décision simple utilisant un seuil fixe et la méthode de décision statistique.

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
7	2	5	17	28,57%	10,53%

FIG. 82 – Résultats de la détection simple, seuil $+/-3$, pour les contenus de type documentaire du corpus de test

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
7	2	5	26	28,57%	7,14%

FIG. 83 – Résultats de la détection simple, seuil $+/-10$, pour les contenus de type documentaire du corpus de test

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
7	3	4	41	42,68%	6,81%

FIG. 84 – Résultats de la détection simple, seuil $+/-21$, pour les contenus de type documentaire de corpus de test

Base	Vérité terrain	Correctes	Manquées	Fausses alarmes	Rappel	Précision
Apprentissage	6	6	0	17	100,00%	36,10%
Test	7	2	5	14	28,57%	12,50%

FIG. 85 – Résultats de la détection statistique pour les contenus de type documentaire du corpus

Les tableaux des figures 82 à 85 montrent les résultats respectifs de la détection simple avec seuil fixé à 3, 10 et 21 et de la détection statistique bayésienne pour les contenus de type série du corpus. De la même manière que dans la section précédente consacrée au résultats des expériences menées sur l'ensemble du corpus tous genres confondus, nous allons maintenant considérer les performances de la détection des frontières de scènes avec les détecteurs audio et vidéo "idéaux". Nous allons aussi calculer les valeurs ajustées de rappel et de précision dans les tableaux des figures ci-dessous.

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
7	5	3	0	3	4	2	2	50,00%

FIG. 86 – Valeur de rappel ajustée sur les contenus de type documentaire du corpus de test avec seuil fixé à $+/-3$

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
7	5	4	0	4	3	1	2	66,66%

FIG. 87 – Valeur de rappel ajustée sur les contenus de type documentaire du corpus de test avec seuil fixé à $+/-10$

Si nous considérons les détecteurs ajustés dans le cas des documentaires, nous remarquons que la règle de décision simple comme la règle de décisions statistique génère moins d'erreurs de classification et que les meilleures performances sont observées pour la règle de décision statistique. Cependant, il faut rester prudent dans nos conclusions car nous n'avions que très

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
7	4	4	0	4	3	0	3	100,00%

Fig. 88 – Valeur de rappel ajustée sur les contenus de type documentaire du corpus de test avec seuil fixé à $+/-21$

Base	R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
Apprentissage	6	0	0	0	0	6	0	6	100,00%
Test	7	5	4	0	4	3	1	2	66,67%

Fig. 89 – Valeur de rappel ajustée sur les contenus de type documentaire du corpus avec la méthode de décision bayésienne

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
7	17	13	4	13	4	2	33,33%

Fig. 90 – Valeur de précision ajustée sur les contenus de type documentaire du corpus de test avec seuil fixé à $+/-3$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
7	26	13	4	13	13	2	13,33%

Fig. 91 – Valeur de précision ajustée sur les contenus de type documentaire du corpus de test avec seuil fixé à $+/-10$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
7	41	13	4	13	38	3	7,31%

Fig. 92 – Valeur de précision ajustée sur les contenus de type documentaire du corpus avec seuil fixé à $+/-21$

Base	P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
Apprentissage	6	17	10	2	10	7	6	46,15%
Test	7	14	15	7	14	0	2	100,00%

Fig. 93 – Valeur de précision ajustée sur les contenus de type documentaire du corpus avec la méthode de décision bayésienne

peu d'échantillons pour nos expérimentations statistiques. Il serait donc nécessaire d'annoter une plus grande quantité de données sur les contenus de type documentaires dans le futur.

Le tableau de la figure 94 résume les résultats des expérimentations menées en considérant les valeurs issues de la vérité terrain pour les détecteurs audio et vidéo. Nous pouvons remarquer que le schéma de décision simple avec seuil est plus performant que la règle de décision statistique en termes de rappel (100% avec un seuil égal à $+/-21$). Cependant, les remarques que nous avons faites ci-dessus concernant la faible quantité d'informations labellisées pour l'entraînement de notre système sont aussi valables pour ces expérimentations.

Règle	Vérité terrain	Correctes	Manquées	Fausses	Rappel	Précision
Jitter+/- 3	7	5	2	4	71,43%	55,56%
Jitter+/- 10	7	6	1	13	85,72%	31,56%
Jitter+/- 21	7	7	0	38	100,00%	15,56%
Statistique (Appr.)	6	6	0	7	100,00%	46,16%
Statistique (Test)	7	6	1	0	85,72%	100,00%

FIG. 94 – Performances idéales, sur les contenus de type documentaire du corpus, du classifieur à partir des valeurs manuelles de la segmentation audio et vidéo

Film

Classe "Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

- Variance (σ_1^2) = 1,00
- Moyenne (μ_1) = 0,50
- Probabilité de classe = *Nombre total de frontières de scènes / Nombre total de frontières de plans de montage* = $27/1174 = 0,02$

Classe "Non-Changement de scène" :

Les paramètres statistiques, relatifs aux valeurs du jitter, obtenus lors de la phase d'apprentissage sont :

- Variance (σ_2^2) = 742930,00
- Moyenne (μ_2) = 80,52
- Probabilité de classe = *Nombre total de non-frontières de scènes / Nombre total de frontières de plans de montage* = $1147/1174 = 0,98$.

Deux méthodes sont comparées : la règle de décision simple utilisant un seuil fixe et la méthode de décision statistique.

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
14	3	11	36	21,43%	7,70%

FIG. 95 – Résultats de la détection simple, seuil $+/-3$, pour les contenus de type film du corpus de test

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
14	4	10	48	28,57%	7,70%

FIG. 96 – Résultats de la détection simple, seuil $+/-10$, pour les contenus de type film du corpus de test

Les tableaux des figures 95 à 98 montrent les résultats respectifs de la détection simple avec seuil fixé à 3, 10 et 21 et de la détection statistique bayésienne pour les contenus de type série du corpus. De la même manière que dans la section précédente consacrée aux résultats des expériences menées sur l'ensemble du corpus tous genres confondus, nous allons maintenant

Vérité terrain	Bonnes détections	Scènes manquées	Fausses alarmes	Rappel	Précision
14	4	10	55	28,57%	6,78%

Fig. 97 – Résultats de la détection simple, seuil $+/-21$, pour les contenus de type film de corpus de test

Base	Vérité terrain	Correctes	Manquées	Fausses alarmes	Rappel	Précision
Apprentissage	13	10	3	40	76,92%	20,00%
Test	14	3	11	32	21,43%	8,57%

Fig. 98 – Résultats de la détection statistique pour les contenus de type série du corpus de test

considérer les performances de la détection des frontières de scènes avec les détecteurs audio et vidéo "idéaux". Nous allons aussi calculer les valeurs ajustées de rappel et de précision dans les tableaux des figures ci-dessous.

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
14	11	7	0	7	7	4	3	42,86%

Fig. 99 – Valeur de rappel ajustée sur les contenus de type film du corpus de test avec seuil fixé à $+/-3$

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
14	10	7	0	7	7	3	4	57,14%

Fig. 100 – Valeur de rappel ajustée sur les contenus de type film du corpus de test avec seuil fixé à $+/-10$

R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
14	10	7	0	7	7	3	4	57,14%

Fig. 101 – Valeur de rappel ajustée sur les contenus de type film du corpus de test avec seuil fixé à $+/-21$

Base	R_A	R_B	R_C	R_D	R_E	R_F	R_G	R_H	R_I
Apprentissage	13	3	3	0	3	10	0	10	100,00%
Test	14	11	9	0	9	5	2	3	60,00%

Fig. 102 – Valeur de rappel ajustée sur les contenus de type film du corpus avec la méthode de décision bayésienne

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
14	36	30	10	27	6	3	33,33%

Fig. 103 – Valeur de précision ajustée sur les contenus de type film du corpus de test avec seuil fixé à $+/-3$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
14	36	30	10	32	18	4	18,18%

FIG. 104 – Valeur de précision ajustée sur les contenus de type film du corpus de test avec seuil fixé à $+/-10$

P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
14	55	30	10	35	25	4	13,80%

FIG. 105 – Valeur de précision ajustée sur les contenus de type film du corpus de test avec seuil fixé à $+/-21$

Base	P_A	P_B	P_C	P_D	P_E	P_F	P_G	P_H
Apprentissage	13	40	35	8	35	5	10	66,67%
Test	14	32	20	11	15	17	3	15,00%

FIG. 106 – Valeur de précision ajustée sur les contenus de type film du corpus avec la méthode de décision bayésienne

Dans le cas des détecteurs "idéaux" (tableaux 99 à 106), la règle de décision simple avec seuil fixé à $+/-3$ donne la meilleure précision et la règle de décision statistique affiche un meilleur taux de rappel. Il semblerait que dans le cas de l'utilisation des détecteurs audio et vidéo automatiques pour les contenus de type film, la règle de décision simple avec une valeur de seuil faible donne les meilleurs résultats.

Règle	Vérité terrain	Correctes	Manquées	Fausses	Rappel	Précision
Jitter+/- 3	14	10	4	6	71,43%	62.50%
Jitter+/- 10	14	11	3	18	78,60%	37,93%
Jitter+/- 21	14	11	3	25	78,60%	30,56%
Statistique (Appr.)	13	13	0	5	100,00%	72,23%
Statistique (Test)	14	12	2	17	85,72%	41,38%

FIG. 107 – Performances idéales, sur les contenus de type film du corpus de test, du classifieur à partir des valeurs manuelles de la segmentation audio et vidéo

Après observation des performances obtenues à l'aide des valeurs issues de la vérité terrain des détecteurs audio et vidéo (tableau 107), nous pouvons remarquer que la meilleure valeur de rappel est obtenue par la méthode statistique. Le meilleur taux de précision est obtenu par la règle de décision simple. Les performances données par la règle de décision bayésienne sont globalement correctes, ici notre méthode de classification statistique rate seulement 2 frontières de scène et génère 17 fausses alarmes. Ces bonnes performances sont dues au fait que nous utilisons les valeurs issues de la vérité terrain des détecteurs audio et vidéo et que le corpus de test relatif aux films est plus conséquent que celui des documentaires.

10.5 Conclusion

Pour conclure, nous avons présenté un vaste panel de résultats afin de valider le modèle de scène audiovisuelle que nous avons proposé et d'estimer les performances réelles de notre méthode statistique de détection des frontières de scène. Nous avons pu remarquer que les performances de la détection statistique sont parfois meilleures que dans le cas d'une décision avec un seuil fixé. Néanmoins, nous devons rester prudents quant à cette affirmation car trop peu de données ont été utilisées pour entraîner nos paramètres statistiques gaussiens (moyenne et variance) et pour tester notre règle de décision bayésienne. Il est nécessaire de valider cette méthode sur un corpus de données de la taille de ceux utilisés dans le cadre de la campagne d'évaluation TRECVID. En dépit de cela, les résultats présentés ici sont de bonne augure pour la suite.

Chapitre 11

Détection des transitions entre les bruits

Dans ce chapitre, nous présentons les résultats de nos expérimentations concernant d'une part la validation de notre ensemble de descripteurs que nous proposons et d'autre part la méthode permettant de caractériser les transitions entre les bruits. C'est ainsi que la première partie de ce chapitre est consacrée aux expérimentations sur les ensembles de descripteurs, tandis que la seconde partie présente les résultats de la méthode de segmentation aveugle.

11.1 Descripteurs proposés

Tout d'abord rappelons quels sont les descripteurs qui composent les deux ensembles que nous souhaitons confronter afin de valider l'ensemble de descripteurs que nous proposons.
Le premier ensemble de descripteurs, qui représente les descripteurs les plus fréquemment utilisés [Pee04], est composé :

- du centroïde spectral,
- du flux spectral,
- du roll-off,
- du taux de passage par zéro et
- des 13 premiers coefficients MFC.

Le second ensemble est composé :

- de la proportion d'énergie dans les bandes de Barks,
- du kurtosis et de sa fréquence,
- du rapport d'auto-corrélation et de sa période et
- du nombre de sinusoïdes [Han03].

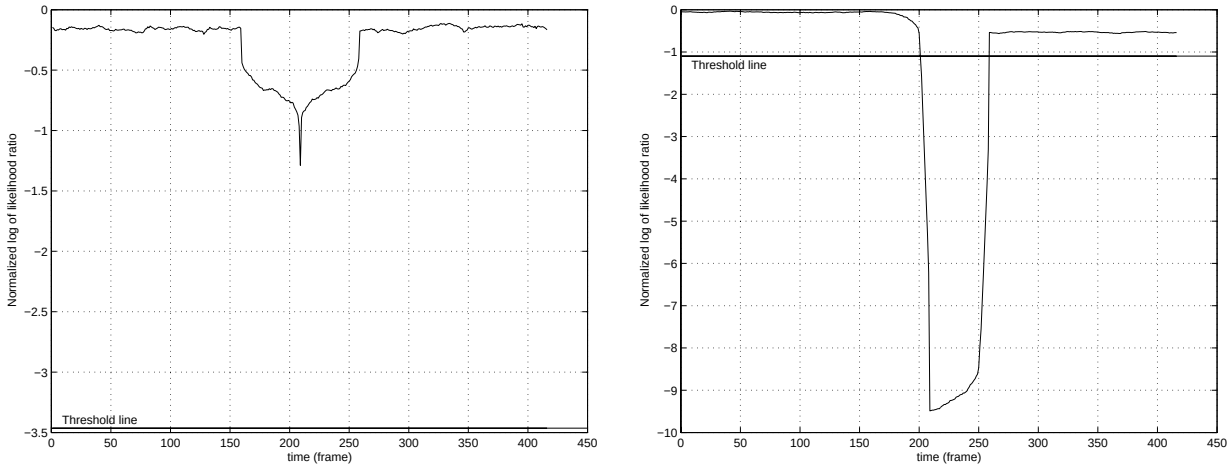
Dans le but de déterminer le relatif pouvoir de discrimination de ces deux ensembles de descripteurs, nous avons réalisé un ensemble d'expérimentations à la fois sur des banques de sons synthétiques et réels. Les bruits réels ont été extraits des bandes sonores de contenus audiovisuels. Le corpus de sons synthétiques est composé de sons bruités impulsifs, pseudo-périodiques, avec sinus et colorés. Nou avons crée un corpus de test synthétique composé de 10

échantillons nous permettant ainsi de caractériser l'ensemble des combinaisons intra et inter classes possible (voir figure 109. Ces sons ont été synthétisés selon la méthode développée dans [Han03] basée sur le modèle CNSS présenté au cours du chapitre 8.

Le rapport de vraisemblance est normalisé en fonction du cardinal de chacun des ensembles de descripteurs décrits ci-dessus. À partir des équations (134) et (135), nous considérons la valeur normalisée de M que nous notons $\tilde{M}$, définie comme :

$$\tilde{M} = \frac{M}{Card(Ensemble\_Descripteurs)} \tag{154}$$

Concernant le protocole expérimental utilisé, nous avons considéré l'ensemble des transitions entre les bruits intra et inter classes (voir figure 109). Puis, nous avons calculé l'ensemble des descripteurs du premier groupe puis du second groupe pour chacun de sons bruites. Après quoi, nous avons soumis les descripteurs de chaque groupe à la méthode de décision statistique et nous avons mesuré l'amplitude de la courbe du maximum de vraissemble au niveau du minimum global qui correspond à une transition. Nous avons considéré la valeur de ce minimum à laquelle nous avons retranché la valeur du minimum local le plus proche et nous considérons l'ensemble en valeur absolue. À titre d'exemple, sur la figure 108, pour le premier groupe nous avons une amplitude de $|(-1,30) - (-0,78)| = 0,52$ et pour le second groue $|(-9.50) - (-0.15)| = 9.35$.
Les résultats de cette comparaison sont données par la figure 108. Ces résultats ont été obtenus à partir des sons du corpus synthétique. La figure 108a montre le logarithme rapport de vraisemblance normalisé pour les descripteurs du premier groupe. Le résultat pour le groupe de descripteurs que nous proposons est fourni par la figure 108b.



a) b)

FIG. 108 – Logarithme du rapport de vraisemblance normalisé pour les groupes de descripteurs : a) Premier groupe, b) Second groupe proposé.

Nous pouvons remarquer que le second groupe de descripteurs possède un pouvoir discriminatif plus important que le premier groupe. Ceci se traduit par une valeur de minimum du logarithme rapport de vraisemblance normalisé dans le cas de la présence d'une transition entre deux types de bruits.

Dans le tableau de la figure 109, nous pouvons aussi remarquer que les différences entre le minimum global et le plus proche minimum, par rapport au minimum global, de la courbe des valeurs du logarithme rapport de vraisemblance normalisé sont supérieures avec les descripteurs du groupe 2. La première colonne du tableau de la figure 109 indique le type de transition testée, et annotée manuellement au préalable. Voici la liste de correspondance des abréviations :

- **i−i** signifie que la transition est de type Impulsif/Impulsif,
- **i−p** signifie que la transition est de type Impulsif/Pseudo-périodique,
- **i−s** signifie que la transition est de type Impulsif/Sinus,
- **i−c** signifie que la transition est de type Impulsif/Coloré,
- **p−s** signifie que la transition est de type Pseudo-périodique/Impulsif,
- **p−p** signifie que la transition est de type Pseudo-périodique/Pseudo-périodique,
- **p−c** signifie que la transition est de type Pseudo-périodique/Coloré,
- **s−s** signifie que la transition est de type Sinus/Sinus,
- **s−c** signifie que la transition est de type Sinus/Coloré et
- **c−c** signifie que la transition est de type Coloré/Coloré.

La seconde colonne précise la différence entre les deux types de bruits, plus précisément il s'agit du paramètre que nous avons fait varier lors de la synthèse des sons bruités considérés. Les transitions entre deux classes sont repérées par le mot-clef "Classe". Enfin, les deux dernières colonnes sont composées de la différence absolue entre le minimum global et le plus proche minimum, par rapport au minimum global, de la courbe des valeurs du logarithme rapport de vraisemblance normalisé.

Transitions	**Différences**	**Gr. 1**	**Gr. 2**
i1−i2	Période	0.50	4.10
i1−i3	Amplitude des impulsions	0.32	0.40
i−p	Classe	0.50	19.00
i−s	Classe	0.30	4.20
i−c	Classe	0.35	3.40
p−s	Classe	1.17	7.80
p1−p2	Amplitude de ACR	10.50	18.65
p1−p3	Période	10.00	18.80
p−c	Classe	0.52	9.35
s1−s2	Amplitude des sinus	1.30	4.85
s1−s3	Nombre de sinus	0.60	2.55
s−c	Classe	0.55	4.50
c1−c2	Couleur	0.68	3.86

Fig. 109 – Comparaison des pouvoirs discriminatifs des deux groupes de descripteurs (i− impulsif, p− pseudo-périodique, c− coloré et s− avec sinus)

Nous remarquons que cette différence admet des valeurs supérieures, en valeur absolue, dans le cas du second groupe. Donc, le pouvoir de discrimination du second groupe de descripteurs est supérieur au premier groupe. Ceci nous permet de conclure que l'ensemble de

descripteurs proposés semble mieux adapté à la caractérisation des bruits. Il est nécessaire de rappeler que les descripteurs que nous proposons ne sont ni nouveaux, ni originaux. Mais c'est leur application au domaine des sons bruités qui est originale.

11.2 Segmentation aveugle

De nombreuses combinaisons entre les bruits ont été réalisées artificiellement. En effet, nous avons créé des échantillons de test en combinant des bruits d'une même classe mais avec des paramètres perceptifs différents. Nous avons, bien évidemment, créé des échantillons de tests avec deux sons bruités de deux classes distinctes. L'ensemble de ces échantillons de tests artificiels sont réalisés à partir de sons bruités synthétiques. Enfin, nous avons complété notre corpus de test avec des transitions de bruits réels qui ont été extraites des bandes sonores de contenus audiovisuels. Nous avons récolté un nombre limité d'échantillons réels, 15 échantillons, à partir des bandes sonores.

Comme il est possible de le remarquer d'après l'équation (133), la prise de décision quant à la présence ou non d'une transition de bruits dépend d'un seuil probabiliste et de la précision du détecteur, le paramètre noté n. De plus, comme nous pouvons l'observer d'après le tableau de la figure 109 les valeurs des minima du logarithme du rapport de vraisemblance normalisé sont supérieures pour le second groupe de descripteurs. Ceci nous permet de conclure que le seuil devrait être adaptatif et nous l'entraînons sur les premières mesures en supposant la continuité du signal du bruit. Avec cette hypothèse, la détection semi-aveugle donne de meilleures performances lorsque nous utilisons le second groupe de descripteurs. Les tableaux des figures ci-dessus en témoignent.

Enfin, les tableaux 110 et 111 illustrent les résultats obtenus par la méthode de segmentation statistique sur le corpus d'échantillons réels en utilisant respectivement le premier et le second groupe de descripteurs. Nous remarquons que les performances de la caractérisation

Vérité terrain	Correctes	Manquées	Fausses	Rappel	Précision
15	8	7	5	53,33%	61,54%

FIG. 110 – Résultats sur des exemples positifs réels avec le premier groupe de descripteurs

Vérité terrain	Correctes	Manquées	Fausses	Rappel	Précision
15	13	2	2	86,67%	86,67%

FIG. 111 – Résultats sur des exemples positifs réels avec le second groupe de descripteurs

statistique des transitions entre les bruits sont meilleures dans le cas de l'utilisation du second groupe de descripteurs. Ceci s'explique par le fait que, comme nous l'avons vu précédemment, le pouvoir de discrimination de la méthode statistique est plus important en utilisant l'ensemble de descripteurs que nous avons proposés. Ainsi, pour conclure, nous pouvons dire que l'ensemble de descripteurs proposés ainsi que la méthode de décision statistique sont bien adaptés au problème posé. Toutefois, ces performances doivent être confirmées par des essais à plus grande échelle.

11.3 Perspectives

Dans ces travaux, nous nous sommes uniquement intéressés à la caractérisation des transitions entre les bruits. Cependant, comme le montre le tableau 112, les frontières de scènes peuvent être caractérisées par d'autres types de transitions dans le flux audio. Les données présentées dans le tableau 112 font suite à une étude menée sur le corpus de test vidéo utilisé dans l'évaluation des performances de la méthode de détection des frontières de scène (voir chapitre 10).

Type de transition audio	Pourcentage
Parole à Bruit	14,28%
Bruit à Bruit	12,24%
Bruit à Parole	12,24%
Musique à Parole	12,24%
Parole à Musique	10,21%
Musique à Bruit	8,16%
Musique à Musique	8,16%
Parole à Parole	6,13%
Pas de transition	6,13%
Musique à (Bruit + Musique)	4.09%
Bruit à (Musique + Parole)	2,04%
(Bruit + Parole) à (Bruit + Parole)	2,04%
Bruit à Musique	2,04%

FIG. 112 – Pourcentage d'occurrences de chaque type de transition audio du corpus vidéo.

Parmi ces différentes transitions, nous avons par exemple des transitions de type **parole à bruit**. Elles sont caractérisées par la présence d'un locuteur (masculin ou féminin) et d'un changement soudain vers du bruit. Dans le domaine vidéo, la transition vidéo associée intervient souvent après la transition audio. Ces types de frontières de scènes peuvent être caractérisées par la mise en place d'une méthode d'analyse de la parole et du bruit, comme par exemple une méthode de classification audio.

Enfin, pour avoir un système de segmentation cross-média complet, il faudrait inclure, sous forme de module, des méthodes qui permettraient de caractériser l'ensemble de ces transitions audio. Cependant, il y a des cas (6,13%) pour lesquels il n'y a pas la présence de transition audio. Une telle situation constitue un cas pathologique dans l'analyse cross-média puisqu'il n'y a aucun moyen de les caractériser. D'autre part, ce pourcentage inclut aussi les cas où la valeur du jitter est excessivement grande pour être associée à une frontière de scène multimédia.

Chapitre 12

Classification des sons bruités

Dans ce chapitre, nous allons présenter les résultats des expériences menées pour valider la méthode de classification des sons bruités que nous avons proposée. Pour cela, nous avons découpé ce chapitre en deux sous-parties dans lesquelles nous allons introduire et développer respectivement le corpus audio utilisé pour l'entraînement et les tests du système de classification ainsi que les expérimentations réalisées et les résultats obtenus.

12.1 Corpus audio

Dans cette partie, nous introduisons la composition des différents corpus que nous avons utilisés. En effet, nous avons été amené à composer deux principales bases de données audio : une base pour l'apprentissage du système de classification statistique et une base pour la validation de la méthode de classification bayésienne proposée. Les deux sous-parties ci-dessous sont consacrées respectivement à la description du contenu des bases de données audio d'apprentissage et de test.

Base d'apprentissage

Dans la section 9.4 de la partie III, nous avons présenté la méthode utilisée pour la phase d'entraînement du système de classification. Nous avons aussi précisé que cet entraînement avait été réalisé à partir d'une base de données audio composée de sons bruités synthétisés. Enfin, nous avons rappelé les grandes lignes de la méthode utilisée [Han03] pour la synthèse de ces sons. Ainsi, dans cette sous-partie, nous détaillons la composition de notre base de données audio d'apprentissage

Pour chaque groupe de sons bruités définis (voir section 9.2) nous avons généré un ensemble de 900 sons échantillonnés à 44100Hz d'une durée de 1 seconde. Cela correspond à 2700 sons sur l'ensemble des trois groupes : impulsif, pseudo-périodique et avec sinusoïdes. Les non-groupes sont composés d'un mélange de sons provenant des deux groupes complémentaires au groupe considéré. À titre d'exemple, le groupe non-impulsif est composé d'un mélange de 450 sons bruités, choisis aléatoirement, provenant du groupe périodique et de 450 sons bruités, choisis aléatoirement, issus du groupe avec sinusoïdes.

Base de test

Dans cette sous-partie, nous nous intéressons à la composition du corpus audio de test utilisé pour la validation du modèle de classification statistique.

Les tests expérimentaux ont été réalisés sur une base de données audio composée de sons bruités naturels. Cette base de données contient 913 échantillons de durée variable, la durée totale étant de 1 heure et 45 minutes. L'ensemble de ces sons bruités a été téléchargé depuis Internet[1]. Il s'agit de sons libres de droit pour l'utilisation et la distribution. Ces sons bruités représentent un large éventail de bruits présents dans la nature. Nous avons unifié le format des données des sons. En effet par rapport au schéma de fonctionnement de notre méthode d'extraction des descripteurs, nous devons convertir l'ensemble des fichiers audio numériques dans un format de données non compressé. Le format utilisé est le format de fichier `WAV` avec les caractéristiques suivantes : 44100kHz pour le taux d'échantillonnage, 16 bits pour l'encodage des échantillons et monophonique. Nous avons opté pour ce corpus car il contient une large variété de sons différents ce qui lui donne sa richesse. Parmi ces sons nous trouvons, par exemple, différentes ambiances (bar, stade de football, ...), des bruits de moteurs, bruits de foule (rire, cris, applaudissements, ...), ...

12.2 Expérimentations et résultats

Dans cette partie, nous présentons le détail des expérimentations menées dans une première sous-partie. Nous exposons les résultats de ces expérimentations dans la seconde et dernière sous-partie.

Expérimentations

Le flux audio sur lequel nous travaillons est découpé en fenêtres d'analyse de deux tailles différentes pour le calcul des descripteurs audio :

- 512 échantillons pour le calcul du rapport d'autocorrélation (ACR) et du nombre de sinus et
- 4096 échantillons pour l'estimation de la fréquence des pulsations.

Dans le cas du calcul du rapport d'autocorrélation (ACR) et du nombre de sinus les fenêtres d'analyse ne se chevauchent pas. En revanche, le calcul de la fréquence des pulsations nécessite un chevauchement de 128 échantillons des fenêtres pour une plus grande précision dans l'estimation. Pour chaque fenêtre d'analyse les descripteurs audio sont calculés et soumis à la méthode de décision statistique. Ainsi, pour chaque fenêtre d'analyse, le système nous retourne un vecteur de trois valeurs correspondant à la décision d'appartenance à chaque groupe de sons bruités défini (voir section 9.2).

Après avoir traité l'ensemble des fenêtres d'analyse que contient le flux audio, nous calculons les proportions de fenêtres étiquetées impulsif, pseudo-périodique et sinus. Ces proportions sont représentées sous forme décimale par une valeur comprise entre 0 et 1. À titre d'exemple, si pour un fichier audio nous obtenons le vecteur des proportions égal à $0,01/0,96/0,02$, alors nous pouvons dire que le son analysé appartient à la classe des sons bruités pseudo-périodiques,

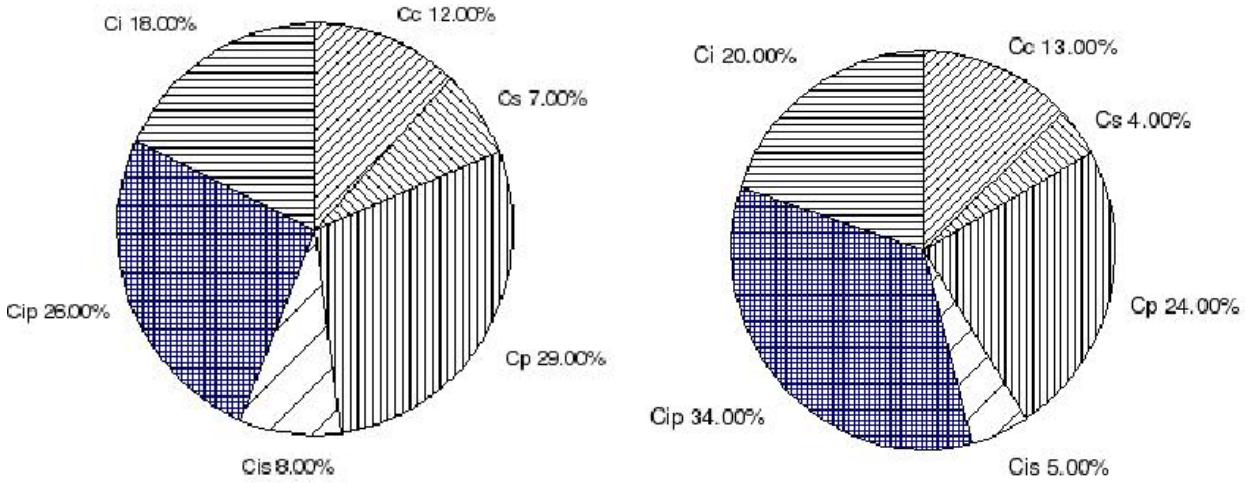
[1] `http ://www.sound-fishing.net/`

notée C_p, car la proportion de fenêtres impulsives, $0,01$ et sinus, $0,02$, est très faible, proche de 0, par rapport à celle de fenêtres pseudo-périodiques, $0,96$ qui est proche de 1.

Résultats

Du point de vue humain, il est très difficile, voire impossible, de conclure avec certitude sur une des classes des sons bruités définies dans la section 9.2. En effet, d'un individu à l'autre l'avis peut diverger du moment que les classes de sons bruités sont caractérisées par des paramètres liés à la perception. Néanmoins, nous avons décidé de réaliser deux types d'expérimentations :

- Tout d'abord, nous regroupons les sons bruités purs ou quasi-purs. Ces sons sont caractérisés par une forte proportion, plus de $0,9$, de fenêtres ayant été classifiées suivant l'une des six classes proposées. Cette répartition est donnée par la figure 113a sous forme de pourcentage. Notons que seul un sous-ensemble des sons testés répondent à cette caractéristique. Ce sous-ensemble est composé de 261 sons, ce qui représente environ 29% des 913 sons de la base de test complète.
- Puis, nous avons regroupé les sons bruités qui sont caractérisés par une proportion de fenêtres, ayant été classifiées suivant l'une des six classes proposées, supérieure à $0,5$. En d'autres termes, nous avons effectué une quantification par seuillage des proportions calculées. Si la valeur dépassé $0,5$, alors nous attribuons 1 à la proportion considérée. Dans le cas contraire, nous affectons 0. Cette répartition est donnée par la figure 113b sous forme de pourcentage. Notons qu'ici, l'ensemble des 913 sons bruités de tests sont pris en considération.



a) b)

Fig. 113 – Proportion des sons bruités classifiés : a) sons bruités appartenant de manière fortement probable à une classe, b) l'ensemble des sons bruités de la base de test.
C_c : Bruits colorés, C_s : Bruits composés de sinusoïdes, C_p : Bruits pseudo-périodiques, C_{is} : Bruits impulsifs et composés de sinusoïdes, C_{ip} : Bruits impulsifs et pseudo-périodiques, C_i : Bruits impulsifs.

Le premier item abordé ci-dessus concerne les sons bruités pour lesquels il est assez facile de déterminer avec certitude leur classe. Cet ensemble est composé de sons bruités tels que

les applaudissements (impulsif), les bruits de moteurs (pseudo-périodique), des bruits de pas (impulsifs), ...

Le premier diagramme circulaire (figure 113a) permet de schématiser la répartition des sons naturels bruités purs ou quasi-purs, en d'autres termes appartenant clairement à une des classes pré-définies. De plus, nous pouvons remarquer que ces répartitions sont bien équilibrées et représentatives du monde environnant qui est composé en grand partie de sons bruités impulsifs, pseudo-périodiques et à la fois impulsifs et pseudo-périodiques. Par ces résultats, nous validons aussi le choix de classes proposées. En effet, aucune de ces classes ne regroupe un fort pourcentage de sons, le pourcentage le plus élevé étant de 29% pour la classe des sons bruités pseudo-périodiques (notée C_p). Tandis que, le pourcentage le plus faible est de 7% pour la classe des sons avec sinus (notée C_s).

Étant entendu que, d'une part, l'estimation des répartitions données par la figure 113a est correcte et que, d'autre part, les deux diagrammes circulaires (figures 113a et 113b) montrent approximativement les mêmes répartitions de classe, nous sommes en mesure de pouvoir affirmer que les performances de la méthode de classification appliquée à l'ensemble complet des sons bruités (voire figure 113b) sont correctes. Les éventuelles erreurs de classification peuvent apparaître dans le cas où l'estimation de la proportion d'une classe se situe au voisinage de $0,5$. En effet, ces situations correspondent souvent à des sons naturels bruités dont le contenu n'est pas homogène. En d'autres termes, l'ensemble des fenêtres d'analyse traitées appartiennent à des classes différentes. Il n'est pas possible d'observer une homogénéité de l'ensemble des fenêtres et de caractériser une classe en particulier.

Pour finir, le tableau de la figure 114 montre les résultats de la classification de quelques exemples de sons bruités parmi l'ensemble de ceux qui constituent la base de test (voir section 12.1). La première colonne indique la dénomination du son bruité naturel considéré. En écoutant ces sons bruités nous pouvons déterminer avec quasi-certitude sa classe, cette classification manuelle est reportée dans la seconde colonne du tableau de la figure 114. Puis, le vecteur des proportions des trois groupes définis, à savoir impulsif, pseudo-périodique et avec sinus, de l'ensemble fenêtres analysées est donné par la troisième colonne. Pour rappel, ce vecteur est composé des proportions de fenêtres impulsives, pseudo-périodiques et avec sinus. De plus, le mot-clé "Imp / Per" signifie que le son bruité étudié appartient à la classe "Impulsif/Périodique" qui correspond à l'intersection des groupes impulsif et pseudo-périodique (voir section 9.2). De même, "Imp / Sinus" correspond à la classe "Impulsif/Sinus".

Son bruité	**Vérité terrain**	**Résultats de classification**
Bruit de pas	Impulsif (C_i)	0.84/0.02/0.00
Insecte volant	Pseudo-Périodique (C_p)	0.00/0.99/0.01
Ambiance de fête	Sinus (C_s)	0.00/0.03/0.96
Rasoir électrique	Imp / Per (C_{ip})	0.51/0.98/0.00
Ronflements humain	Imp / Per (C_{ip})	0.98/0.97/0.00
Applaudissements	Impulsif (C_i)	0.98/0.06/0.00
Bruit de carillon	Imp / Sinus (C_{is})	0.99/0.09/0.90
Fantômes	Pseudo-Périodique (C_p)	0.00/0.98/0.02
Tempête	Coloré (C_c)	0.00/0.09/0.00

FIG. 114 – Résultats de la classification pour quelques exemples de sons naturels bruités.

12.3 Application de la classification : la segmentation semi-aveugle

Une des applications de la classification des sons bruités est la segmentation semi-aveugle (cf section 8.3). Dans ce cas, nous parlons de segmentation semi-aveugle car nous avons connaissance à priori de la classe de bruit considérée. Cette information nous permet de sélectionner manuellement des descripteurs qui caractérisent cette classe au mieux et de ne pas considérer les autres descripteurs. Ainsi, nous tentons de caractériser les transitions entre les bruits d'une même classe, nous qualifions ces transitions d' *intra-classe*.

Nous avons mené des expérimentations semblables à celles présentées dans la chapitre 11 précédent. En effet, nous avons produit manuellement des sons bruités contenant des transitions intra-classe. Pour cela, nous avons concaténé des sons bruités *synthétiques* entre eux. Nous avons procédé de la même manière avec des sons *naturels* considérés comme appartenant à la même classe au regard des résultats obtenus à l'aide de la méthode de classification automatique présentée dans la section 9.3.

La tableau de la figure 115 présente les différents types de transitions que nous avons produites. La première colonne de ce tableau indique la nature du son bruité considéré : soit naturel, soit synthétique. La seconde colonne renseigne sur le type de la transition considérée : soit pseudo-périodique, soit impulsive, soit avec sinus, soit coloré. Enfin la troisième et dernière colonne précise le descripteur sélectionné en fonction du type de la transition.

Nature du son bruité	Type de transition	Descripteur utilisé
Synthétique	Impulsive	Kurtosis
Synthétique	Pseudo-périodique	Rapport d'autocorrélation (ACR)
Synthétique	Avec Sinus	Nombre de sinus
Synthétique	Coloré	Énergie dans les bandes de barks
Naturel	Impulsive	Kurtosis
Naturel	Impulsive	Kurtosis
Naturel	Pseudo-périodiques	ACR
Naturel	Pseudo-périodiques	ACR
Naturel	Avec Sinus	Nombre de sinus
Naturel	Coloré	Énergie dans les bandes de barks

Fig. 115 – Détail des sons bruités générés pour le test de la méthode de segmentation semi-aveugle

Bien qu'ayant été menés sur un corpus restreint, les résultats obtenus sont prometteurs. En effet, le tableau de la figure 116 montre les performances observées en fonction de la nature des sons bruités considérés.

Nature du bruit	Vérité terrain	Transitions ratées	Fausses alarmes	Rappel	Précision
Naturel	6	1	4	83,33%	55,56%
Synthétique	4	0	1	100 %	80%
Naturel + Synthétique	10	1	5	90%	64%

Fig. 116 – Performances de la méthode de segmentation aveugle

Nous observons des performances moindres dans la cas de sons naturels. Ceci est, en partie, lié au phénomène de propagation de l'erreur. En effet, les performances de la méthode de classification biaisent les performances finales de la segmentation semi-aveugle. Nous avons rencontré le même cas pour la détection des frontières de scènes qui est liée aux résultats des analyses des flux multimédias (voir chapitre 10).

12.4 Conclusion

Dans cette partie, nous avons proposé une méthode statistique de classification des sons bruités naturels. Ceci en vue de l'application à la segmentation semi-aveugle des bruits afin de caractériser les transitions de bruits au sein d'une même classe.

Bien que la validation des résultats de la classification n'est pas entièrement robuste du fait que nous sommes face à un problème de perception, les résultats obtenus sont prometteurs. D'autant plus que de tels travaux ne sont pas encore très développés dans la littérature. Quant aux résultats de la segmentation semi-aveugle, ils sont faussés par les performances de la classification mais restent, malgré tout, convenables.

Évidemment des tests sur des corpus de grandeur importante restent à mener en perspective.

Conclusion générale et perspectives

Ainsi, dans ce travail de thèse nous avons abordé le difficile problème de la segmentation des documents audiovisuels en scènes. Nous avons tenté de proposer une méthode suffisamment générique, convenable à un vaste ensemble de contenus *artistiques*, sans modèles de prédiction à priori. De l'ensemble des travaux présentés dans cette thèse, nous avons proposé trois grands points :

- un modèle de scène générique adapté aux contenus multimédia artistiques,
- l'analyse et le développement des méthodes de segmentation et de classification des sons bruités et
- l'application de la méthode de classification des sons bruités naturels : la segmentation semi-aveugle.

Enfin, il est important de préciser que pour l'ensemble des méthodes statistiques développées dans le cadre de cette thèse, nous sommes restés dans le cadre de la décision bayésienne.
En conclusion, nous pouvons dire que nous avons juste effleuré le vaste domaine d'indexation vidéo dans lequel l'indexation cross-média est sans aucun doute la voie la plus prometteuse.

Dans la première partie, nous avons proposé un modèle générique de frontières de scènes en s'intéressant à la corrélation temporelle entre une frontière de plan de montage vidéo et les frontières d'un silence audio. Nous avons proposé et développé à la fois, une approche de décision statistique basée sur le schéma bayésien et une mesure temporelle entre une transition vidéo et la transition audio la plus proche en termes de nombre d'images vidéo. Pour réaliser l'ensemble du système, nous avons utilisé un détecteur de frontière de plan de montage et un détecteur de silence déjà développés que nous avons légèrement adaptés à notre problème.
Enfin, bien que ne couvrant pas l'ensemble des frontières de scène, notre modèle est adapté à plus de la moitié des cas. De plus, ce modèle se veut générique, sans a priori ni sur le type de contenu audiovisuel, ni sur le type de scène. Toutefois, ce modèle a aussi ses limites. Il peut arriver que nous soyons en présence d'une coïncidence temporelle entre un silence et une transition vidéo à l'intérieur d'une scène, donc d'une frontière de scène selon notre modèle. Ce cas est typique des scènes de dialogue dans les contenu de type série. C'est pourquoi, l'utilisation d'un détecteur de scènes de dialogue permettrait d'améliorer les performances globales du système.

C'est la raison pour laquelle, dans un second temps, nous nous sommes intéressés à la caractérisation des transitions entre les bruits présents dans le flux audio. Ces transitions, selon notre étude sur un vaste corpus de contenus artistiques télédiffusés, sont parmi les plus fréquentes sur les frontières de scènes. Dans ces travaux de caractérisation des transitions de type bruit/bruit, nous avons tout d'abord proposé un ensemble de descripteurs adaptés au traitement des sons bruités. Nous avons, aussi, proposé une méthode de segmentation aveugle

pour la détection des changements de bruits. Enfin, nous avons montré que, pour notre problème, l'utilisation des descripteurs audio permet à la méthode de segmentation statistique d'accroître ses performances aussi bien dans le cas de sons bruités naturels que de sons bruités synthétiques.
Dans les expériences, nous avons montré que l'ensemble des descripteurs audio que nous proposons est pertinent pour l'étude des sons bruités. En effet, la méthode de segmentation semi-aveugle offre de meilleures performances qu'elle ne le fait avec les descripteurs classiques. Avec un ensemble de descripteurs classiques nous obtenons 53,33% de rappel et 61,54% de précision sur un ensemble de test réel. Tandis que dans le mêmes conditions mais avec l'ensemble de descripteurs proposés nous obtenons 86,67% à la fois pour le rappel et la précision. Toutefois, des tests à plus grande échelle doivent être menés afin de confirmer ces premiers résultats prometteurs.

Dans le troisième partie de ces travaux de recherche, nous avons proposé une méthode statistique de classification des sons bruités. Pour cela, nous avons défini six classes de sons bruités caractérisées par trois descripteurs audio. Ces descripteurs audio ont été choisis car ils permettent de caractériser les sons bruités au niveau de la perception : perception de la hauteur, perception des impulsions et perception de sinusoïdes. Les algorithmes de calcul de ces descripteurs audio ont été développés et expérimentés sur des sons bruités naturels. Les résultats des expérimentations confirment le bien fondé du choix des classes et de leur descripteur associé, car la répartition des sons bruités selon l'ensemble des classes est assez représentatif de la réalité (répartition non-uniforme). Enfin, nous avons aussi montré que la sélection d'un descripteur spécifique à une classe permet de caractériser les transitions entre deux bruits qui appartiennent à une même classe. Nous avons utilisé les résultats fournis par la méthode de classification des sons bruités pour déterminer la classe et donc sélectionner le descripteur adéquat.

Après avoir résumé l'ensemble des travaux menés, nous présentons leurs perspectives dans les paragraphes qui suivent.

Dans le cadre de nos travaux sur la segmentation des bruits nous nous sommes aussi intéressés par une étude plus approfondie au sujet des descripteurs audio. Le but étant de mettre au point un nouvel ensemble de descripteurs enrichis afin d'améliorer encore les performances de segmentation des sons bruités.

En ce qui concerne la méthode de classification des sons bruités que nous avons proposée, nous pensons que certaines classes pourraient être subdivisées et ainsi affiner les résultats de classification. À titre d'exemple, les pourcentages de sons bruités des classes C_p et C_{ip} (resp. périodique et impulsif/périodique) sont importants comparés aux autres. C'est la raison pour laquelle, le groupe pseudo-périodique pourrait être subdivisé en prenant en compte le nombre de hauteurs perçues, par exemple.

Enfin, les applications de cette approche de classification sont nombreuses. La première étant d'améliorer la caractérisation des transitions de type bruit/bruit en pondérant les descripteurs associés à la classe considérée. En effet, cela permettrait de sélectionner un ou plusieurs descripteurs adéquats en fonction du type de bruit analysé. Finalement, cette classification de bruits a des applications plus larges dans l'ensemble de la problématique d'indexation. La fusion adéquate des descripteurs des sons bruités et des descripteurs vidéo doit permettre de résoudre le problème d'indexation des scènes d'intérieur et d'extérieur dans les contenus audiovisuels car l'indexation des événements est une autre ouverture de ces approches.

Annexe A

Éléments de traitement du signal

Dans cette annexe, nous introduisons quelques notions concernant le traitement des signaux dans les domaines spectral et temporel. Ces notions sont très largement utilisées dans le domaine de l'analyse audio.

A.1 Fonction de convolution

La technique de convolution s'applique à deux signaux et permet des transformations sonores [Roa97]. L'opération de convolution, notée $*$, de deux signaux x et y est définie par les équations suivantes, respectivement dans le cas de signaux continus et discrets :

$$\forall t,\ \ x(t) * y(t) = \int_{-\infty}^{+\infty} x(k)y(k-t)dk \tag{155}$$

$$\forall n,\ \ x[n] * y[n] = \sum_{k=-\infty}^{+\infty} x[k]y[k-n] \tag{156}$$

A.2 Fonction d'autocorrélation

L'étude des périodicités d'un signal peut s'effectuer à partir de l'observation des similitudes du signal avec ce même signal décalé temporellement. La fonction d'autocorrélation permet d'étudier ces similitudes. Elle peut être considérée dans deux cas, selon que le signal soit de durée finie ou infinie. Nous ne traiterons ici que du cas fini, la fonction d'autocorrélation est alors définie comme étant la corrélation d'un signal avec lui-même décalé, d'où :

$$\forall \tau,\ \Gamma(\tau) = \int_{-\infty}^{+\infty} x(t)x(t+\tau)dt \tag{157}$$

avec τ le temps de décalage et Γ la fonction d'autocorrélation.

Cette fonction d'autocorrélation est en général appliquée à des signaux discrets de longueur finie pour la détection de périodicités. Dans ce cas, l'équation (157) devient :

$$\forall \tau \in [0, W_s - k], \Gamma(\tau) = \frac{1}{W_s} \sum_{i=0}^{k} x[i]x[i+\tau] \tag{158}$$

avec W_s le nombre total d'échantillons du signal x et k un paramètre entier de sorte que la périodicité recherchée doit notamment être inférieure à la durée correspondant au nombre d'échantillons $W_s - k$.

Enfin, la fonction d'autocorrélation admet les propriétés suivantes :

- Il s'agit d'une fonction paire :
$$\forall \tau, \Gamma(\tau) = \Gamma(-\tau)$$
- Comme la fonction d'autocorrélation traduit la similitude entre le signal et une version retardée de lui-même, il est évident qu'elle prend une valeur maximale lorsque le retard est nul ($\tau = 0$) :
$$\forall \tau, \Gamma(0) \geq |\Gamma(\tau)|$$
Seul le cas de signaux périodiques permet de définir des valeurs de la fonction égales aux valeurs de l'origine ($\Gamma(0) = \Gamma(\tau), \forall \tau \neq 0$). Dans ce cas précis, la fonction d'autocorrélation est périodique. Ce maximum est l'énergie totale du signal, puisqu'il est défini par l'intégrale de la puissance du signal :
$$E_0 = \Gamma(0) = \int_{-\infty}^{+\infty} x^2(t)dt$$
- Dans le cas de signaux non infinis, qui exclut le cas de signaux périodiques, la fonction d'autocorrélation converge vers 0 à l'infini :
$$\lim_{\tau \to \infty} |\Gamma(\tau)| = 0$$

A.3 Transformée de Fourier

La transformée de Fourier doit son nom au mathématicien français J.B.Fourier (1768-1830). Elle définit une série d'opérations mathématiques qui permet d'associer à une onde (un signal), une série de sinusoïdes de fréquences, d'amplitudes et de phases déterminées. Elle permet donc de passer de la représentation temporelle (amplitude en fonction du temps) à la représentation spectrale (amplitude en fonction de la fréquence).
Dans le cas d'un signal continu, la transformée de Fourier est définie par :

$$FT(x) = X(f) = \int_{-\infty}^{+\infty} x(t) exp^{2\pi f t} dt \tag{159}$$

avec $X(f)$ le spectre du signal $x(t)$.
La transformée de Fourier possède les propriétés suivantes :

- Il s'agit d'une application linéaire. Ainsi, pour deux signaux x et y, et pour deux réels α et β :
$$FT(\alpha x + \beta y) = \alpha X(f) + \beta Y(f)$$
avec $X(f)$ et $Y(f)$ deux transformées de Fourier.
- La transformée de Fourier d'un produit de deux signaux est la convolution des transformées de Fourier des signaux. Ainsi, pour deux signaux x et y nous avons :
$$FT(xy) = X(f) * Y(f)$$
et
$$FT(x * y) = X(f)Y(f)$$

La transformée de Fourier admet une inverse qui permet de passer de la représentation fréquentielle d'un signal à sa représentation temporelle. Cette opération est définie par :

$$x(t) = \int_{-\infty}^{+\infty} X(f) exp^{2\pi ft} df \tag{160}$$

avec $x(t)$ le signal et $X(f)$ son spectre.
De même que pour la fonction d'autocorrélation, pour être exploitable sur les signaux discrets la transformée de Fourier admet une forme discrète définie par :

$$X(k) = \frac{1}{W_s} \sum_{n=0}^{W_s-1} x(n) e^{-j\frac{2\pi}{W_s}nk} \tag{161}$$

avec x le signal comprenant W_s échantillons de spectre $X(k)$.
Le signal analysé étant discret, la transformée discrète de Fourier définit des amplitudes correspondant à des fréquences discrètes, F_k. Cette discrétisation est liée à la fréquence d'échantillonnage, F_s, mais aussi à la taille de la fenêtre du signal, W_s, pour laquelle le calcul est effectué. Les fréquences f_k ont donc pour valeur :

$$\forall k \in [0, W_s], \quad f_k = k\frac{F_s}{W_s} \tag{162}$$

De la même manière, les amplitudes et les phases sont directement accessibles à partir des complexes, $X(k)$, donnés par la transformée de Fourier :

$$A_k = |X(k)| = \sqrt{(Re(X(k)))^2 + (Im(X(k)))^2} \tag{163}$$

$$\theta_k = arg(X(k)) = arctan\frac{Im(X(k))}{Re(X(k))} \tag{164}$$

Nous terminerons ce paragraphe sur la transformée de Fourier par la présentation de la transformation rapide de Fourier. En effet, l'algorithme de la transformée de Fourier rapide permet de calculer la transformée de Fourier avec une complexité réduite en $(W_s log_2(W_s))$, W_s, le nombre total d'échantillons du signal, doit être une puissance de 2. De plus amples informations sur cet algorithme sont présentes dans la littérature [Moo90]. De plus, une implémentation de cet algorithme très couramment utilisée est la FFTW (Fatest Fourier Tranform in the West) [FJ98].

Annexe B

Outils statistiques pour l'analyse et l'indexation multimédia

Dans cette annexe, nous introduisons quelques outils statistiques couramment utilisés dans les domaines de l'analyse et l'indexation audio et vidéo.

B.1 Distance Kullback Leibler

La divergence de Kullback-Leibler (KL) entre deux distributions de probabilité x et y est aussi connue sous le nom d'entropie relative :

$$KL(x,y) = \sum_{j=1}^{n} x_j log\left(\frac{x_j}{y_j}\right) \tag{165}$$

Comme $KL(x,y) \neq KL(y,x)$, il ne s'agit pas réellement d'une distance. C'est pourquoi, il est possible de définir la distance de Kullback-Leibler (DKL) (1951) comme :

$$DKL(x,y) = DKL(y,x) = KL(x,y) + KL(y,x) \tag{166}$$

De plus, lorsque les deux distributions sont des gaussiennes, on peut montrer que :

$$\Delta = \frac{(\sigma_1^2 - \sigma_2^2)^2 + (\mu_1 - \mu_2)^2(\sigma_1^2 + \sigma_2^2)}{4\sigma_1^2\sigma_2^2} \tag{167}$$

Enfin, cette distance est considérée comme étant la meilleure mesure pour la recherche d'informations dans les grandes bases de données [MM02].

B.2 Probabilités conditionnelles et théorème de Bayes

La notion de probabilité conditionnelle est une des notions les plus fructueuses de la théorie des probabilités : elle s'intéresse à la probabilité d'un événement A dans le cadre de répétitions d'une expérience aléatoire où l'on ne retient que les cas où un événement B est réalisé (il est

nécessaire que B ne soit pas impossible et soit réalisé au cours de l'expérience) et s'introduit naturellement dans une interprétation *fréquentiste* des probabilités.

En effet, si N est le nombre de répétitions d'une expérience aléatoire, N_E le nombre de réalisations de l'événement E au cours de cette série de répétitions, E pouvant être l'événement A, B, ou $A \cap B$, alors :

- la fréquence d'occurrence de A est $\frac{N_A}{N}$.
- la fréquence d'occurrence de B est $\frac{N_B}{N}$.
- la fréquence d'occurrence de $A \cap B$ est $\frac{N_{A\cap B}}{N}$.

Si l'on ne s'intéresse qu'aux répétitions de l'expérience ayant réalisé B, N_B est le nombre de répétitions de ces expériences et la fréquence d'occurrence de A sur ce type de répétitions est le rapport (si $N_B \neq 0$) :

$$\frac{N_{A\cap B}}{N_B} = \frac{\frac{N_{A\cap B}}{N}}{\frac{N_B}{N}} \tag{168}$$

La probabilité de A conditionnelle en B (probabilité de A sachant B) s'interprète comme la limite de ce rapport quand N tend vers $+\infty$. On comprend dès lors les définitions suivantes.

B.2.1 Définitions

Soit un espace de probabilité $\{\Omega, \mathcal{A}, P\}$, la probabilité conditionnelle $P(A|B)$ d'un évènement A conditionnelle à l'événement B tel que $P(B) > 0$ est définie par :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{169}$$

Si $P(B) = 0$ cette définition n'a aucun sens, si $A \subseteq B$ alors $P(A|B) = \frac{P(A)}{P(B)}$ et on a $P(B|B) = 1$.

De même, si l'on se place dans le cas d'un espace d'épreuves Ω fini (et non vide $Card(\Omega) > 0$) et d'une mesure de probabilité uniforme sur Ω on retrouve la même intuition des probabilités conditionnelles issue du *calcul des chances*. Si $Card(B) > 0$, on peut écrire directement :

$$P(A) = \frac{Card(A)}{Card(\Omega)} \tag{170}$$

$$P(B) = \frac{Card(B)}{Card(\Omega)} \tag{171}$$

$$P(A \cap B) = \frac{Card(A \cap B)}{Card(\Omega)} \tag{172}$$

$$P(A|B) = \frac{Card(A \cap B)}{Card(B)} \tag{173}$$

La définition ci-dessus n'a aucun sens si $P(B) = 0$ à cause de son occurrence au dénominateur du quotient de l'expression 169. Cela correspond tout à fait à l'intuition (*fréquentiste* ou non) que l'on peut donner des probabilités conditionnelles : un événement impossible n'est jamais réalisé dans l'expérience, donc $0 \leq P(A \cap B) \leq P(B) = 0$ et on ne peut pas évaluer quoi que ce soit sur la probabilité de A dans les cas où on a aussi B puisque cela n'arrive jamais.

Cette interprétation est cependant tout aussi valable pour les *subjectivistes* en ce qu'un événement de probabilité nulle peut être vu comme impossible : les probabilités conditionnelles sont d'ailleurs majoritairement utilisées par les tenants des approches probabilistes *subjectivistes* [Pea88, Tri72].

Il est possible de réécrire le quotient de l'expression 169 de la manière suivante :

$$P(A \cap B) = P(A|B).P(B) \tag{174}$$

Cependant cette écriture ne généralise en rien la définition liée à l'équation 169 et ne fournit pas plus une définition de $P(A|B)$ dans le cas où $P(B) = 0$.

La probabilité de A sachant B (ou conditionnelle en B) est la probabilité que l'on aurait trouvé pour A si l'espace d'épreuves était limité aux cas réalisant B au lieu de l'ensemble Ω. En particulier la fonction d'ensemble P_B qui à tout événement $A \in \mathcal{A}$ associe $P_b(A) = P(A|B)$ est également une mesure de probabilité sur $(\Omega, \mathcal{A})$, mais son support est inclus dans B ($\forall A \in \mathcal{A}, A \cap B = \varnothing$). On peut donc définir l'espérance conditionnelle d'une variable aléatoire X à valeur dans un espace E (voir section B.2.3).

B.2.2 Théorème de Bayes

Le théorème de Bayes est une conséquence immédiate de la loi de composition des probabilités qui est un des axiomes fondamentaux de la théorie des probabilités.

Si A et B sont deux événements, cette loi de composition des probabilités indique que la probabilité $P(AB)$ d'observer à la fois A et B est simplement donnée par :

$$P(A \cap B) = P(AB) = P(A) \times P(B|A) = P(B) \times P(A|B) \tag{175}$$

Le théorème de Bayes découle directement de l'équation (175) précédente, d'où :

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{176}$$

avec $P(A|B)$ la probabilité *a posteriori* de A sachant B, $P(B|A)$ la probabilité *a priori* de A sachant B, $P(A)$ et $P(B)$ les probabilités *a priori* des événements A et B respectivement. Les probabilités $P(A)$ et $P(B)$ sont aussi appelées *probabilités marginales*, et $P(A|B)$ la *fonction de vraisemblance* de A pour un B connu.

Enfin, ce théorème peut se généraliser sans peine au cas de plusieurs événements.

B.2.3 Espérance conditionnelle

La probabilité conditionnelle $P(A|B)$ peut être vue comme la probabilité de A lorsque B devient l'événement certain à la suite d'une information affirmant que B est vérifié ($P(B) = 1$). On remplace donc la mesure de probabilité P par la mesure P_B sur $(\Omega, \mathcal{A})$. Cette démarche est utilisée pour *réviser* la mesure de probabilité P sachant l'information B dans les approches bayésiennes.

B.3 Notion de vraisemblance

Étant donné un échantillon observé $(x_1, \ldots, x_n)$ et une loi de probabilité P_θ, la vraisemblance quantifie la probabilité que les observations proviennent effectivement d'un échantillon (théorique) de la loi P_θ.

Considérons $C = \{c_1, \ldots, c_k\}$ un ensemble fini, $\{P_\theta\}$ une famille de lois de probabilité sur C, et n un entier naturel inférieur ou égal à k. On appelle vraisemblance associée à la famille $\{P_\theta\}$, la fonction qui à un n-uplet $(x_1, \ldots, x_n)$ d'éléments de C et à une valeur θ du paramètre associe la quantité :

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} P_\theta(x_i) \tag{177}$$

L'interprétation est la suivante. Considérons un ensemble de variables aléatoires $(X_1, \ldots, X_n)$ de la loi P_θ. Par définition, les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi P_θ. Donc la probabilité que l'ensemble $(X_1, \ldots, X_n)$ ait pour réalisation l'échantillon observé $(x_1, \ldots, x_n)$ est le produit des probabilités pour que X_i prenne la valeur x_i, à savoir :

$$\mathbb{P}[(X_1, \ldots, X_n) = (x_1, \ldots, x_n)] = L(x_1, \ldots, x_n, \theta) \tag{178}$$

Dans le cas d'un modèle continu sur $\mathbb{R}$, la loi P_θ a une densité sur $\mathbb{R}$, et la probabilité pour que l'échantillon prenne une valeur particulière est toujours nulle. Il faut alors remplacer la probabilité P_θ par sa densité dans la définition de la vraisemblance.

Maintenant, prenons $\{P_\theta\}$ une famille de lois de probabilité continues sur $\mathbb{R}$ et n un entier. Notons f_θ la densité de probabilité de la loi P_θ. On appelle vraisemblance associée à la famille $\{P_\theta\}$, la fonction qui à un n-uplet $(x_1, \ldots, x_n)$ d'éléments de C et à une valeur θ du paramètre associe la quantité :

$$L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f_\theta(x_i) \tag{179}$$

L'interprétation est la suivante. Considérons un ensemble de variables aléatoires $(X_1, \ldots, X_n)$ de la loi continue P_θ. Soit ε un réel strictement positif (petit). La probabilité que l'échantillon théorique $(X_1, \ldots, X_n)$ ait une réalisation proche *à ε près* de l'échantillon observé $(x_1, \ldots, x_n)$ peut s'écrire :

$$\begin{array}{rcl} \mathbb{P}\left[X_1 \in [x_1 - \frac{\varepsilon}{2}, x_1 + \frac{\varepsilon}{2} \text{ et } \ldots \text{ et } X_n \in [x_n - \frac{\varepsilon}{2}, x_n + \frac{\varepsilon}{2}\right] & = & \prod_{i=1}^{n} \int_{x_i - \frac{\varepsilon}{2}}^{x_i + \frac{\varepsilon}{2}} f_\Theta(x) dx \\ & \simeq & \prod_{i=1}^{n} f_\Theta(x_i) dx \\ & = & \varepsilon L(x_1, \ldots, x_n, \Theta) \end{array} \tag{180}$$

Estimer un paramètre par la méthode du maximum de vraisemblance, c'est proposer comme valeur de ce paramètre celle qui rend maximale la vraisemblance, à savoir la probabilité d'observer les données comme réalisation d'un échantillon de la loi P_θ.

Supposons, maintenant, que pour toute valeur $(x_1, \ldots, x_n)$, la fonction qui à θ associe $L(x_1, \ldots, x_n, \theta)$ admette un maximum unique. La valeur $\widehat{\theta}$ pour laquelle ce maximum est atteint dépend de $(x_1, \ldots, x_n)$:

$$\widehat{\theta} = \tau(x_1, \ldots, x_n) = \arg\max L(x_1, \ldots, x_n, \theta) \tag{181}$$

On appelle cela le principe d'estimation par maximum de vraisemblance.

Si $(X_1, \ldots, X_n)$ est un échantillon (théorique) de la loi P_θ, la variable aléatoire :

$$T = \tau(X_1, \ldots, X_n) \tag{182}$$

est l'estimateur du maximum de vraisemblance de θ.

Pour la plupart des lois de probabilité usuelles, l'estimateur du maximum de vraisemblance est défini de façon unique, et se calcule explicitement. Sur le plan théorique, il présente de nombreux avantages. Sous des hypothèses vérifiées par de nombreux modèles courants, on démontre qu'il est asymptotiquement sans biais et convergent. On démontre de plus que sa variance est minimale. La méthode du maximum de vraisemblance est donc théoriquement la meilleure des méthodes d'estimation. C'est ce que nous allons présenter dans la section suivante.

B.3.1 Principe du maximum de vraisemblance

Dans la plupart des cas d'intérêt pratique, la loi P_θ, et donc aussi la vraisemblance, ont une expression dérivable par rapport à θ. Pour calculer le maximum de la vraisemblance, il faut déterminer les valeurs pour lesquelles la dérivée de la vraisemblance s'annule. Or par définition, la vraisemblance est un produit de probabilités ou de densités, qui peut être assez compliqué à dériver. Il est préférable de dériver une somme, et c'est pourquoi on commence par remplacer la vraisemblance par son logarithme. La fonction logarithme étant croissante, il est équivalent de maximiser $\log(L(x_1, \ldots, x_n, \theta))$ ou $L(x_1, \ldots, x_n, \theta)$. Une fois déterminée une valeur de θ pour laquelle la dérivée s'annule, il faut s'assurer à l'aide de la dérivée seconde que ce point est bien un maximum. Nous traitons ci-dessous le cas d'une distribution Normale.

Nous présentons en annexe (voir annexe B.4) le développement du principe du maximum de vraisemblance dans le cas d'une loi Normale.

Si $(X_1, \ldots, X_n)$ est un échantillon de la loi normale de paramètres μ et σ, les estimateurs du maximum de vraisemblance de μ et σ sont respectivement la moyenne et la variance empirique de l'échantillon (cf. Annexe B.4 pour cette démonstration).

B.4 Principe du maximum de vraisemblance dans le cas d'une distribution Normale

Pour un paramètre multidimensionnel, le principe est le même, mais les calculs d'optimisation sont plus compliqués. Pour les lois normales, deux paramètres sont inconnus. Pour un n-uplet de réels $(x_1, \ldots, x_n)$ la vraisemblance vaut :

$$L(x_1, \ldots, x_n, \mu, \sigma 2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2} \tag{183}$$

Son logarithme est :

$$\log(L(x_1, \ldots, x_n, \mu, \sigma^2)) = -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=0}^{n}(x_i - \mu)^2 \tag{184}$$

Les dérivées partielles par rapport aux paramètres μ et σ sont :

$$\frac{\partial \log(L(x_1,\ldots,x_n,\mu,\sigma^2))}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=0}^{n}(x_i-\mu) \tag{185}$$

et

$$\frac{\partial \log(L(x_1,\ldots,x_n,\mu,\sigma^2))}{\partial (\sigma^2)} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=0}^{n}(x_i-\mu)^2 \tag{186}$$

Elle s'annulent pour :

$$\widehat{\mu} = \frac{\sum x_i}{n} \text{ et } \widehat{\sigma^2} = \frac{\sum (x_i-\widehat{\mu})^2}{n} \tag{187}$$

Les dérivées partielles secondes valent :

$$\frac{\partial^2 \log(L(x_1,\ldots,x_n,\mu,\sigma^2))}{\partial \mu^2} = -\frac{n}{\sigma^2} \tag{188}$$

$$\log\left(L(x_1,\ldots,x_n,\mu,\sigma^2)\right)\partial\mu\partial(\sigma^2) = -\frac{2}{\sigma^3}\sum(x_i-\mu) \tag{189}$$

$$\frac{\partial^2 \log(L(x_1,\ldots,x_n,\mu,\sigma^2))}{\partial (\sigma^2)^2} = -\frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum(x_i-\mu)^2 \tag{190}$$

La matrice hessienne (matrice des dérivées partielles secondes) au point $(\widehat{\mu},\widehat{\sigma^2})$ est donc :

$$\begin{pmatrix} -\frac{n}{\widehat{\sigma^2}} & 0 \\ 0 & -\frac{n}{\widehat{\sigma^2}} \end{pmatrix}$$

Elle est définie négative, le point $(\widehat{\mu},\widehat{\sigma^2})$ est bien un maximum.

Annexe C

Détecteur de silence pour le domaine compressé

Le flux audio compressé est au format MPEG-1 couche 2, la méthode que nous proposons est basée sur le calcul de l'énergie du spectre du signal à partir des coefficients d'échelle. Ces coefficients d'échelle sont convertis en exposants, les exposants étant à la base du codage AC-3. Dans le paragraphe suivant, nous présentons ce que sont les coefficients d'échelle et les exposants.

La méthode présentée ci-dessous est le fruit de travaux menés au sein de Philips Research Natlab (NL) ce qui implique qu'elle est soumise aux droits légaux de propriété intellectuelle (brevet) [SN]. De plus, dans cette thèse nous n'avons fait qu'utiliser cette méthode sans apporter de modifications.

C.1 Coefficients d'échelle et exposants

Le noyau du codeur/décodeur (codec) MPEG-1 couche 2 [BS94] consiste en la décomposition d'un signal numérique non compressé en sous-bandes à l'aide d'une banque de filtres séparateurs de bandes. Plus précisément, le filtre séparateur de bandes est un jeu de filtres à phase linéaire, ayant tous la même largeur de bande et qui se recouvrent. La sortie de chaque bande consiste en des échantillons représentatifs de la forme d'onde. Dans chaque bande de fréquence, l'entrée audio est amplifiée au maximum avant la transmission.
En MPEG, le nombre de filtres est de 32. L'axe du temps est divisé en blocs d'égale longueur. A l'intérieur de chaque bande, le niveau est amplifié par multiplication jusqu'à sa valeur maximale. Le gain nécessaire est constant pour la durée du bloc et un seul **facteur d'échelle** est transmis avec chaque bloc pour chaque bande de façon à pouvoir renverser le processus au décodage. La sortie du groupe de filtres est également analysée afin de déterminer le spectre du signal d'entrée. Cette analyse permet de réaliser un modèle de masquage permettant de déterminer le degré de masquage que l'on peut attendre dans chaque bande. Dans chaque bande, plus le masquage est agissant, moins l'échantillon doit être précis. La précision d'échantillon est alors réduite par re-quantification en vue de diminuer la longueur des mots. Cette réduction est aussi constante pour chaque mot dans la bande, mais les différentes bandes peuvent utiliser des longueurs de mot différentes. La longueur de mots doit être transmise comme un

code d'affectation de bits afin de permettre au décodeur de dé-sérialiser convenablement le flux de bits.

Le système AC-3 [TDD+94, Jea] est basé sur la transformée de Fourier du signal et l'on obtient le gain de codage en re-quantifiant les coefficients de fréquence. L'entrée PCM d'un codeur AC-3 est divisée en blocs par des fenêtres qui se chevauchent. Ces blocs contiennent chacun 512 échantillons mais, du fait du chevauchement total, il existe une redondance de 100%. La figure 117 illustre ce chevauchement.
Après la transformée, il existe donc 512 coefficients qui peuvent, du fait de la redondance,

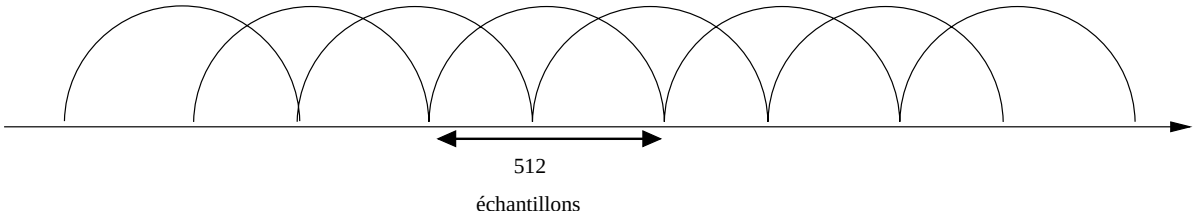


FIG. 117 – Illustration du principe de chevauchement pour le codage audio AC-3

être ramenés à 256 à l'aide d'une technique appelée Suppression par aliasing dans le domaine temporel (TDAC, Time Domain Aliasing Cancelation). La forme du signal d'entrée est analysée et, s'il existe une évolution significative dans la seconde moitié du bloc, le signal sera séparé en deux pour éviter les pré-échos. Dans ce cas, le nombre de coefficients reste le même mais la résolution de fréquence sera divisée par deux et la résolution temporelle doublée. Un indicateur(flag) est placé dans le flux de bits pour signaler que cette opération a été effectuée. Les coefficients sont émis sous un format à virgule flottante avec une mantisse et un exposant. La représentation est l'équivalent binaire de la notation scientifique.
Les **exposants** constituent en fait les **facteurs d'échelle**. Le jeu d'exposants d'un bloc produisent l'analyse spectrale d'un signal d'entrée avec une précision finie sur une échelle logarithmique appelée enveloppe spectrale. Cette analyse spectrale est le signal d'entrée du modèle de masquage définissant, pour chaque fréquence, le niveau jusqu'où le bruit peut être augmenté. Le modèle de masquage pilote le processus de re-quantification qui diminue la précision de chaque coefficient en arrondissant la mantisse. Cette mantisse constitue une partie significative de la donnée transmise. Les exposants sont également transmis mais pas intégralement dans la mesure où la redondance qu'ils comportent peut être ultérieurement exploitée.
A l'intérieur d'un bloc, seul le premier **exposant**(celui de la fréquence la plus basse) est transmis dans sa forme absolue. Les autres sont transmis de façon différentielle et le décodeur ajoute la différence avec l'**exposant** précédent. Quand le signal audio présente un spectre assez aplati, les exposants peuvent être identiques pour plusieurs bandes de fréquences. Les exposants peuvent alors être assemblés en groupes de deux à quatre avec un indicateur décrivant leur mode de groupement. Des jeux de six blocs sont assemblés dans une trame de synchro AC-3. Le premier bloc de la trame comporte la donnée complète pour l'exposant mais, dans le cas de signaux constants, les blocs suivants de la trame peuvent utiliser le même **exposant**.

C.2 Description de l'algorithme

L'algorithme utilisé peut se décomposer en quatre grandes étapes :

1. Conversion des facteurs d'échelles en exposants,
2. Calcul de l'énergie du spectre,
3. Mise à jour de la moyenne de l'énergie du spectre et
4. Détection des possibles silences.

C.2.1 Conversion des facteurs d'échelles en exposants

La première étape consiste en la conversion des facteurs d'échelle en exposants. D'après l'analyse des codages MPEG-1 et AC-3, nous pouvons mettre en correspondance les facteurs d'échelle avec les exposants de la manière suivante :

- Prendre la valeur en dB (décibels) du facteur d'échelle à convertir (multiplier par -2 la valeur du facteur d'échelle),
- Diviser le résultat par -5.8 dB et
- Prendre l'entier le plus proche du résultat de la division précédente.

Le tableau de la figure 118 donne la correspondance entre les deux termes.

Facteur d'échelle	Exposant
0,1,2	0
3,4,5	1
6,7,8	2
9,10,11	3
12,13	4
14,15,16	5
17,18,19	6
20,21,22	7
23,24,25	8
26,27,28	9
29,30,31	10
32,33,34	11
35,36,37	12
38,39,40	13
41,42	14
43,44,45	15
46,47,48	16
49,50,51	17
52,53,54	18
55,56,57	19
58	20

Fig. 118 – Table de correspondance entre les facteurs d'échelle et les exposants

C.2.2 Calcul de l'énergie du spectre

La deuxième phase de l'algorithme consiste à calculer l'énergie locale d'une fenêtre du spectre. Cette énergie est obtenue grâce à la formule suivante :

$$S(K) = \sum_{i=0}^{E-1} \left(2^{-exp_{K,i}}\right)^2 \tag{191}$$

avec $exp_{K,i}$ le i-ième exposant dans la fenêtre d'analyse d'indice K et E le nombre d'exposants utilisés pour le calcul de l'énergie locale. En d'autres termes, E correspond à la taille de la fenêtre d'analyse. Pour pouvoir appliquer cette formule, il faut tout d'abord convertir les facteurs d'échelle de chaque fenêtre d'analyse en exposants en utilisant le tableau de correspondance de la figure 118.

C.2.3 Mise à jour de la moyenne de l'énergie du spectre

La troisième phase consiste à mettre à jour la moyenne de l'énergie du spectre. Rappelons que la moyenne de l'énergie du spectre du signal est calculée à partir des énergies locales d'un ensemble de M fenêtres d'analyse consécutives, incluant la fenêtre K comme étant la dernière de l'ensemble. Les expérimentations ont montré que de considérer le logarithme des valeurs d'énergie du spectre plutôt que les valeurs elles-mêmes donnent un pouvoir de discrimination supérieur pour la détection des silences. La moyenne de l'énergie du spectre, $S_A(K)$ est donnée par la formule suivante :

$$S_A(K) = \frac{1}{M} \sum_{i=0}^{M-1} log(S(K-i)) \tag{192}$$

C.2.4 Détection des possibles silences

La dernière phase de l'algorithme consiste à mettre en place une règle de décision pour la détection des frontières des silences. Il existe des méthodes de décision statistiques qui pourraient permettre de résoudre ce problème. Toutefois, notre objectif étant de mettre en place un processus temps réel nous nous limiterons à une règle de décision basée sur l'utilisation d'un seuil. Donc, une fenêtre d'analyse audio est étiquetée silence si la condition suivante est respectée :

$$\frac{log(S(K))}{S_A(K)} \leq T_{compressed} \tag{193}$$

avec $T_{compressed}$ un seuil adaptatif pour le domaine compressé obtenu de manière expérimentale en fonction du genre vidéo étudié.

Une solution alternative pourrait consister en une comparaison entre l'énergie locale du spectre avec une estimation du niveau de bruit. Nous estimons de manière rapide le niveau du bruit de fond en recherchant la valeur minimale de l'énergie du spectre calculée du début du signal jusqu'à la fenêtre d'analyse courante. Néanmoins, la règle de décision donnée par l'équation (193) produit de meilleures performances, c'est la raison pour laquelle nous avons opté pour cette solution.

Bibliographie

[AAD+03] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C.-Y. Lin, A. Natsev, M. Naphade, C. Neti, H. J. Nock, H. H. Permuter, R. Singh, J. R. Smith, S. Srinivasan, B. L. Tseng, T. V. Ashwin, and D. Zhang. Ibm research trec-2002 video retrieval system. In *Proc. of TRECVID Workshop'03*, 2003.

[AAW01] A. A. Alatan, A. N. Akansu, and W. Wolf. Multi-modal dialog scene detection using hidden markov model for content-based multimedia indexing. *Multimedia Tools and Applications*, 14 :137–151, 2001.

[ABL01] S. Allegro, M. Büchler, and S. Launer. Automatic sound classification inspired by auditory scene analysis. In *Proc. of Eurospeech*, Aalborg, Denmark, 2001.

[AJ94] P. Aigrain and P. Joly. The automatic real-time analysis of film editing and transition effects and its applications. 18(1) :93–103, 1994.

[And03] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Willey Series in Probability and Statistics, 2003.

[AO98] R. André-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(1) :29–40, 1998.

[AR76] B. S. Atal and L. R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24(3) :201–212, 1976.

[Bas88] M. Basseville. Detecting changes in signals and systems - a survey. *Automatica*, 24 :309–326, 1988.

[BCC99] F. Beritelli, S. Casale, and A. Cavallaro. A multi-channel speech/silence detector based on time delay estimation and fuzzy classification. In *Proc. ICASSP'99*, volume 1, pages 93–96, 1999.

[BG96] P. Bouthemy and F. Ganansia. Video partitioning and camera motion characterization for content-based video indexing. In *Proc. of 3rd IEEE ICIP*, volume 1, pages 905–909, 1996.

[BGG99] P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Trans. on Circuits and Systems for Video Technology*, 9(7) :1030–1044, 1999.

[BGR+99] P. Bouthemy, C. Garcia, R. Ronfard, G. Tziritas, E. Veneau, and D. Zugaj. Scene segmentation and image feature extraction for video indexing and retrieval. In *Proc. of VISUAL'99*, volume 1614, pages 245–252, Amsterdam, 1999.

[BP02] E. Bruno and D. Pellerin. Video shot detection based on linear prediction of motion. In *Proc. of ICME'02*, 2002.

[BPDCL02] J. Benois-Pineau, M. Desainte-Catherine, and N. Louis. A method for extraction of audio-visual leitmotif in movies by cross media analysis. In EURASIP, editor, *EUSIPCO'2002*, volume III, pages 345–348, 2002.

[BPJ02] C. Burges, J. Platt, and S. Jana. Extracting noise-robust features from audio data. In *Proc. of the ICASSP'02*, 2002.

[BS94] K. Brandenburg and G. Stoll. Iso-mpeg-1 audio : A generic standard for coding of high quality digital audio. *Journal of the Audio Engineering Society*, 42(10) :780–792, 1994.

[BSP99] G. Boccignone, M. De Santo, and G. Percannella. Joint audio-video processing of mpeg encoded sequences. In *Proc. of ICMCS'99*, volume 2, pages 225–229, 1999.

[Cas] *Page Internet concernant le project CASSANDRA chez Philips Research.* http ://www.extra.research.philips.com/cassandra.

[CBPBM99] F. Coudert, J. Benois-Pineau, D. Barba, and E. Malan. Fast motion-based content extraction for video indexing. In *Proc. of CBMI'99*, pages 123–129, Toulouse, 1999.

[CBPLB99] F. Coudert, J. Benois-Pineau, P.-Y. Le Lann, and D. Barba. Binkey : A system for video content analysis "on the fly". In *Proc. of ICMCS'99*, pages 679–684, Florence, 1999.

[CCK+03] L. Chaisorn, T.-S. Chua, C.-K. Koh, Y. Zhao, H. Xu, H. Feng, and Q. Tian. A two-level multi-modal approach for story segmentation of large news video corpus. In *Proc. of TRECVID Workshop'03*, 2003.

[CH78] D.R. Cox and D.V. Hinkley. *Theoritical statistics.* London, Chapman and Hall, 1978.

[CSLZ02] S.-C. Chen, M.-L. Shyu, W. Liao, and C. Zhang. Scene change detection by audio and video clues. In *Proc. of ICME'02*, pages 365–368, 2002.

[CSZK01] S.-C. Chen, M.-L. Shyu, C.-C. Zhang, and R. L. Kashyap. Video scene change detection method using unsupervised segmentation and object tracking. In *Proc. of ICME'01*, 2001.

[DBAF99] A. Dufaux, L. Besacier, M. Ansorge, and F.Pellandini. Automatic classification of wideband acoustic signals. *Joint 137th meeting of the Acoustical Society of America and Forum Acousticum 99*, pages 14–19, 1999.

[DBAP00] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini. Automatic sound detection and recognition for noisy environment. In *Proc. of EUSIPCO'00*, pages 1033–1036, Tampere, Finland, 2000.

[DBP01] M. Durik and J. Benois-Pineau. Robust motion characterisation for video indexing based on mpeg-2 optical flow. In *Proc. of CBMI'01*, pages 57–64, 2001.

[DCBP05] A. Don, L. Carminati, and J. Benois-Pineau. Detection of visual dialog scenes in video content based on structural and semantic features. In *Proc. of CBMI'05*, June 2005.

[Del00] P. Delacourt. *La segmentation et le regroupement par locuteurs pour l'indexation de documents audio*. Thèse de doctorat, Institut Eurécom, 2000.

[DHS00] R. Duda, P. Hart, and D. Stork. Pattern classification. *John Wiley & Sons*, 2000.

[DW99] P. Dellacourt and C. Wellekens. Audio data indexing : Use of second-order statistics for speaker-based segmentation. In *IEEE ICMCS'99*, Florence, Italy, 1999.

[DW00] P. Delacourt and C.J. Wellekens. Distbic : A speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1-2) :111–126, 2000.

[EB04] A. Eleftheriadis and P. Batra. Optimal data partitioning of mpeg-2 coded video. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(10) :1195–1209, 2004.

[EMKPK00] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal. Speech/music discrimination for multimedia applications. In *Proc of ICASSP'00*, pages 2445–2448, 2000.

[EMSK99] K. El-Maleh, A. Samouelian, and P. Kabal. Frame level noise classification in mobile environments, 1999.

[ENRC03] T. En-Najjary, O. Rosec, and T. Chonavel. Influence de la modélisation spectrale sur les performances d'un système de conversion de voix. In *Proc. of GRETSI 2003*, Septembre 2003.

[Ero01] A. Eronen. Automatic musical instrument recognition. Master's thesis, Tampere University of Technology, 2001.

[FH52] E. Fix and J. L. Hodges. Discriminatory analysis : nonparametric discrimination : small sample performance. *Technical Report No. 11. Project No. 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas*, 1952.

[FJ98] M. Frigo and S.G. Johnson. Fftw user's manual. 1998.

[Foo97] J. T. Foote. Content-based retrieval of music and audio. *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, 3229 :138–147, 1997.

[GLA01] J.-L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communication*, 2001.

[Goo96] M. Goodwin. Residual modeling in music analysis-synthesis. In *Proceedings of the IEEE International Conference On Acoustics, Speech, and Signal Processing (ICASSP'96), Atlanta, GA, USA*, pages 1005–1008, 1996.

[GPD00] F. Gouyon, F. Pachet, and O. Delerue. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proc. of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, 2000.

[GR02] G.Peeters and X. Rodet. Automatically selecting signal descriptors for sound classification. In *Proc. of ICMC'02*, Goteborg, Denmark, 2002.

[Ha03] A. Hauptman and al. Informedia at trec 2003 : Analyzing and searching broadcast news video. In *Proceedings of TRECVID 2003*, November 2003.

[Han02] A. Hanjalic. Shot-boundary detection : Unraveled and resolved ? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2) :90–105, 2002.

[Han03] P. Hanna. Modélisation statistique de sons bruités : étude de la densité spectrale, analyse, transformation musicale et synthèse. *PhD thesis, LaBRI, Université Bordeaux 1*, 2003.

[Har97] W.M. Hartmann. Signals, sound and sensation. *Modern Acoustics and Signal Processing AIP Press*, 1997.

[HBDC02] P. Hanna, A. Beurivé, and M. Desainte-Catherine. Real-time noise synthesis with control of the spectral density. *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFX'02), Hamburg, Germany*, pages 151–156, 2002.

[HC02] H. Harb and L. Chen. Technique de classification du signal sonore en des classes sonores. *Pending Patent nf 02 08 548*, Juillet 2002.

[HC03a] H. Harb and L. Chen. Cyndi : Un moteur d'indexation de la bande sonore par une segmentation sémantique et extraction de mots clés. In *Proc. of CORESA'03*, 2003.

[HC03b] H. Harb and L. Chen. Gender identification using a general audio classifier. In *Proc. of ICME'03*, Baltimore, USA, 2003.

[HCA01] H. Harb, L. Chen, and J.-Y. Auloge. Speech/music/silence and gender detection algorithm. In *Proc. of the 7th International Conference on Distributed Multimedia Systems (DMS'01)*, 2001.

[HClK+03] W. Hsu, S.-F. Chang, l. Kennnedy, C.-W. Huang, C.-Y. Lin, and G. Iyengar. Discovery and fusion of salient multi-modal features towards news story segmentation. In *Proc. of TRECVID Workshop 2003*, 2003.

[HDC03] P. Hanna and M. Desainte-Catherine. Using statistical analysis of the intensity fluctuations to detect sinusoids in noisy signals. Technical report, LaBRI, University of Bordeaux 1, 2003. http ://www.labri.fr/Labri/Publications/Rapports-internes/.

[HDC04] P. Hanna and M. Desainte-Catherine. Cnss model : a statistical and spectral model for representing noisy sounds with short-time sinusoids. Technical report, LaBRI, University of Bordeaux 1, 2004. http ://www.labri.fr/Labri/Publications/Rapports-internes/.

[HDC05] P. HANNA and M. DESAINTE-CATHERINE. A statistical and spectral model for representing noisy sounds with short-time sinusoids. *EURASIP JASP 2005*, 12(2005) :1794–1806, 2005.

[HLDCBP04] P. Hanna, N. Louis, M. Desainte-Catherine, and J. Benois-Pineau. Audio features for noisy sounds segmentation. In *ISMIR 2004*, volume 1, pages 120–124, 2004.

[HMGB86] W.M. Hartmann, S. McAdams, A. Gerzso, and P. Boulez. Discrimination of spectral density. *Journal of Acoustical Society of America*, 79(6) :1915–1925, 1986.

[HS02] H. Harb and L. Shen. Video scene description : An audio based approach. In *Proc. of the 1st MEDIANET 2003 Conference*, pages 243–254, Souss, Tunis, 2002.

[HS03] H. Harb and L. Shen. Highlights detection in sports videos based on audio analysis. In *Proc. of CBMI'03*, IRISA, Rennes, France, 2003.

[HW94] J. D. Hoyt and H. Wechsler. Detection of human speech in structured noise. In *Proc. of the ICASSP'94*, 1994.

[Jea] *Page internet personnelle de Daniel Jean (codage audio).* http ://pages.videotron.ca/danjean/.

[JEA99] S. Jacobs, A. Eleftheriadis, and D. Anastassiou. Silence detection for multimedia communication systems. *Multimedia Systems*, 7(2) :157–164, 1999.

[JYL00] S.-B. Jun, K. Yoon, and H.-Y. Lee. Dissolve transition detection algorithm using spatio-temporal distribution of mpeg macro-block types. In *Proc. of the 8th ACM Int. Conf. on Multimedia*, pages 391–394, Marina Del Rey, California, USA, 2000.

[KCL00] Y.-M. Kim, S. W. Choi, and S.-W. Lee. Fast scene change detection using direct feature extraction from mpeg compressed videos. In *Proc. of the 15th ICPR'00*, 2000.

[KD97] V. Kobla and D. Doermann. Extraction of features for indexing mpeg-compressed video. In *IEEE First Workshop on Multimedia Signal Processing*, pages 337–342, 1997.

[KGG+03] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot. Hmm based structuring of tennis videos using visual and audio cues. In *Proc. of ICME'03*, 2003.

[KGOG03] E. Kijak, G. Gravier, L. Oisel, and P. Gros. Audiovisual integration for tennis broadcast structuring. In *Proc. of CBMI'03*, IRISA, Rennes, France, 2003.

[KW96] D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. In *Proc of Interface Conf.*, 1996.

[Lev98] S. Levine. *Audio Representations for Data Compression and Compressed Domain Processing.* PhD thesis, CCRMA, Stanford University, 1998.

[LJC87] J. Lynch, J. Josemhauns, and R. Crochiere. Speech / silence segmentation for real-time coding via rule based adaptive endpoint detection. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (IEEE)*, pages 1348–1351, 1987.

[LJZ01] L. Lu, H. Jiang, and HJ. Zhang. A robust audio classification and segmentation method. In *Proc. of the 9th ACM International Conference on Multimedia*, Ottawa, Canada, 2001.

[LKC00] Y. Lin, M. S. Kanhanhalli, and T.-S. Chua. Temporal multi-resolution analysis for video segmentation. In *Proc. of SPIE (Storage and Retrieval for Media Databases)*, volume 3972, pages 494–505, Jan 2000.

[LM00] R. Leonardi and P. Migliorati. Semantic indexing of multimedia documents. *IEEE Multimedia*, 9(2) :44–51, 2000.

[LMP04] R. Leonardi, P. Migliorati, and M. Prandini. Semantic indexing of soccer audio-visual sequences : A multimodal approach based on controlled markov chains. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(5) :634–643, 2004.

[Log00] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the first International Symposium on Music Information Retrieval (ISMIR'00)*, Plymouth, Massachusetts, October 2000.

[LPR03] A.A. Livshin, G. Peeters, and X. Rodet. Studies and improvements in automatic classification of musical sound samples. In *Proc. of ICMC'03*, 2003.

[LRRW81] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29(4) :777–785, 1981.

[LSL97] G. Lupatini, C. Saraceno, and R. Leonardi. Scene break detection : A comparison. In *Proc. of RIDE'98*, pages 34–41, 1997.

[LWC98] Z. Liu, Y. Wang, and T. Chen. Audio feature extraction and analysis for scene segmentation and classification. *The Journal of VLSI Signal Processing*, 20(1/2) :61–79, 1998.

[LZJ02] L. Lu, H.-J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *IEEE Transactions On Speech and Audio Processing*, 10(7) :504–516, 2002.

[MAC01] W. Mahdi, M. Ardebilian, and L. Chen. Automatic scene segmentation based on spatial-temporal clues and rhythm. *International Journal of Networking and Informations Systems*, 5, September 2001.

[Mak95] J. Makhoul. Linear prediction : A tutorial review. In *Proc. IEEE*, volume 5, pages 561–580, 1995.

[Mar98] Martin. Musical instrument identification : A pattern-recognition approach. In *Proc. of the 136th meeting of the Acoustical Society of America*, October 1998.

[MB95] A.V. McCree and T.P. Barnwell. Mixed excitation lpc vocoder model for low bit rate speech coding. *IEEE Transaction on Speech and Audio Processing*, 3(4) :242–250, 1995.

[MB96] P. Masri and A. Bateman. Improved modelling of attack transients in music analysis-resynthesis. *Proceedings of International Computer Music Conference (ICMC'96), Hong-Kong, China*, pages 100–103, 1996.

[MB03] M.F. McKinney and J. Breebaart. Features for audio and music classification. In *Proc. of ISMIR'03*, 2003.

[MF03] J. Mas and G. Fernandez. Video shot boundary detection based on color histogram. In *Proc. of TRECVID Workshop'03*, 2003.

[MHI+02] A. Miene, T. Hermes, G. Ioannidis, R. Fathi, and O. Herzog. Automatic shot boundary detection and classification of indoor and outdoor scenes. In *Proc. of TRECVID Workshop 2002*, 2002.

[MM02] B.S. Manjunath and W.-Y. Ma. *Texture features for image retrieval.* 2002.

[Moo90] F.R. Moore. Elements for computer music. *Prentice Hall*, 1990.

[Mor04] D. Moraru. *Segmentation en locuteurs de documents audios et audiovisuels : application à la recherche d'information multimédia.* Thèse de doctorat, Institut National Polytechnique de Grenoble, 2004.

[Mpe] *MPEG-7 : Context, Objectives and Technical Roadmaps (V.12).* ISO/IECJTC1/SC29/WG11/N2861.

[NAT98] J. Nam, M. Alghoniemy, and A. Tewfik. Audio-visual content-based violent scene characterization. In *Proc. of ICIP'98*, volume 1, pages 353–357, 1998.

[NEP+05] J. Nesvadba, F. Ernst, J. Perhavc, J. Benois-Pineau, and L. Primaux. Comparison of shot boundary detectors. pages 788–791, 2005.

[Ney81] H. Ney. An optimization algorithm for determining the endpoints of isolated utterances. In *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing (ICASSP)*, pages 720–723, 1981.

[NK97] Y. Najamura and T. Kanade. Semantic analysis for video contents extraction - spotting by association in news video. In *Proc. of 1997 ACM Int. Conf. on Multimedia*, pages 393–401, 1997.

[NLBP+04] J. Nesvadba, N. Louis, J. Benois-Pineau, M. Desainte-Catherine, and M.K. Middelink. Low-level cross-media statistical approach for semantic partitioning of audio-visual content in home multimedia environment. In *Proc. of IWSSIP'04*, 2004.

[NMBP+04] H. Nicolas, A. Manoury, J. Benois-Pineau, W. Dupuy, and D. Barba. Grouping video shots into scenes based on 1d mosaic descriptors. In *Proceedings of ICIP*, pages 637–639, 2004.

[NPZ02] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Transactions on Multimedia*, 4(4) :446–458, 2002.

[OB95] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4) :348–365, 1995.

[Pau02] S. Pauws. Cubyhum : A fully operational query by humming system. In *Proc. of ISMIR'02*, 2002.

[PBPKD04] L. Primaux, J. Benois-Pineau, P. Krämer, and J-P. Domenger. Shot boundary detection in the framework of rough indexing paradigm. 2004.

[PC96] E. Parris and M.J. Carrey. Language independent gender identification. In *Proc. of the ICASSP'96*, 1996.

[Pea88] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[Pee04] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. *CUIDADO Project Report*, 2004.

[PFE96] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. *Praktische Informatik IV, Univ. Mannheim, Germany*, 1996.

[Pie02] A . Pienimäki. Indexing music databases using automatic extraction frequent phrases. 2002.

[PR98] G. Peeters and X. Rodet. Analyse et synthèse de sons musicaux par la méthode psola. *Proceedings of the Journées d'Informatique Musicale (JIM'98), La Londe-les-Maures, France*, 1998.

[PR99] G. Peeters and X. Rodet. Sinola : A new analysis/synthesis method using spectrum peak shape distortion, phase and reassigned spectrum. *Proceedings of International Computer Music Conference (ICMC'99), Beijing, China*, pages 153–156, 1999.

[PR01] M. J. Pickering and S. M. Rüger. Multi-timescale video shot-change detection. In *Proc. of 10th Text REtrieval Conference (TREC-10)*, 2001.

[Pre00] D. Pressnitzer. Modèles psychoacoustiques et perception de hauteur. In *Proc. of JIM*, 2000.

[PRMAO03] J. Pinquier, J.-L. Rouas, J. Mauclair, and R. André-Obrecht. Détection de la parole et de la musique dans mes documents sonores : fusion de deux approches. In *GRETSI'03*, 2003.

[PSAO02a] J. Pinquier, C. Sénac, and R. André-Obrecht. Indexation de la bande sonore : Recherche des composantes parole et musique. In *Proc. of RFIA'02*, Angers, France, 2002.

[PSAO02b] J. Pinquier, C. Senac, and R. Andreé-Obrecht. Speech and music classification in audio documents. In *ICSLP'02*, 2002.

[PT04] C. Panagiotakis and G. Tziritas. A speech/music discriminator based on rms and zero-crossings. *IEEE Transactions on Multimedia*, 6(5) :676–686, 2004.

[RJ93] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing. Prentice Hall, Englewood Cliffs, NJ, 1993.

[Roa97] C. Roads. Sound transformation by convolution. *Musical Signal Processing*, pages 411–438, 1997.

[Roa98] C. Roads. L'audionumérique (trad. the computer music tutorial). *Dunod*, 1998.

[RR95] D.A. Reynolds and R.C. Rose. Robust test-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3(1), 1995.

[RS75] L. R. Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. *Bell Systems Technical Journal*, Vol. 54 :297–315, February 1975.

[RS77] L. R. Rabiner and M. R. Sambur. Application of an lpc distance measure to the voiced-unvoiced-silence detection problem. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-25 :338–343, 1977.

[RS78] L. Rabiner and R.W. Schafer. Digital processing of speech signals. *Englewood Cliffs, N.J. : Prentice Hall*, 1978.

[RS03] Z. Rasheed and M. Sahah. Scene detection in hollywood movies and tv shows. In *Proc. of Conference on Vision and Pattern Recognition (CVPR'03)*, 2003.

[SA99] J.O. Smith and J.S. Abel. Bark and erb bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7(6) :697–708, November 1999.

[Sau96] J. Saunders. Real-time discrimination of broadcast speech/music. In *Proc. of Int. Conf. Acoustics, Speech, Signal processing'96*, volume 2, pages 993–996, 1996.

[SBY04] J. Song, S.Y. Bae, and K. Yoon. Mid-level melody representation of polyphonic audio for query-by-humming system. In *Proc. of ISMIR'04*, 2004.

[SC00] H. Sundaram and S.-F. Chang. Video scene segmentation using video and audio features. In *Proc. of ICME'00*, 2000.

[She97] B. Shen. Hdh based compressed video cut detection. In *Proc. of VISUAL'97*, San Diego, California, USA, 1997.

[SHM+03] M. Sugano, K. Hoashi, K. Matsumoto, F. Sugaya, and Y. Nakajima. Shot boundary determination on mpeg compressed domain and story segmentation experiments for trecvid 2003. In *Proc. of TRECVID Workshop'03*, 2003.

[SJBS97] M. Siegler, U. Jain, B.Raj, and R. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. of the DARPA Speech Recognition Workshop, The Westfields Conference Center*, 1997.

[SL97] C. Saraceno and R. Leonardi. Audio as a support to scene change detection and characterization of video sequences. In *Proc. of ICASSP'97*, volume IV, pages 2597–2600, Munich, Germany, 1997.

[SN] A. Stella and J. Nesvadba. Silence detection, estimating signal power in compressed audio. WO 02093801. http ://v3.espacenet.com/origdoc ?DB=EPODOC&IDX=WO02093801&F=8 &RPN=WO02093801&DOC=deb45b02b97229e9a36219d19953faca71.

[Sou83] P. De Souza. A statistical approach to the design of an adaptive self-normalizing silence detector. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-31(3) :678–684, 1983.

[SS90] X. Serra and J. Smith. Spectral modeling synthesis : a sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4) :12–24, 1990.

[SS97] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeatures speech/music discriminator. In *Proc. of the ICASSP'97*, 1997.

[SS01] P. Salembier and J. R. Smith. Mpeg-7 multimedia description schemes. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(6) :748–759, 2001.

[Tal95] Talkin. *A Robust Algorithm for Pitch Tracking.* Klein and Paliwal (eds.), 1995.

[TC99] G. Tzanetakis and P. Cook. Multifeature audio segmentation for browsing and annotation. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, 1999.

[TC02] G. Tzanetakis and P. Cook. Musical genre classification of audio signal. *IEEE Transactions on Speech and Audio Processing*, 10(5) :293–302, 2002.

[TC04] G. Tzanetakis and M.-Y. Chen. Building audio classifiers for broadcast news retrieval. In *Proc. of WIAMIS'04*, Lisbona, Portugal, 2004.

[Tch96] B.P. Tchistiakov. *Lectures in probability theory.* In Russian, 1996.

[TDD+94] C. Todd, A. Davidson, F. Davis, D. Fielder, D. Link, and S. Vernon. Ac-3 : Flexible perceptual coding for audio transmission and storage. In *AES 96th convention*, 1994.

[TP99] S. Tsekeridou and I. Pitas. Audio-visual content analysis for content-based video indexing. In *Proc. of ICMCS'99*, volume I, pages 667–672, Florence, Italy, 1999.

[Tri72] M. Tribus. *Décisions rationelles dans l'incertain.* Masson, 1972.

[TSKR00] Y.-P. Tan, D. D. Saur, S. R. Kulkarni, and P. J. Ramadge. Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1) :133–146, 2000.

[TZ04] W. Tavanapong and J. Zhou. Shot clustering techniques for story browsing. *IEEE Transactions on Multimedia*, 6(4) :517–527, 2004.

[Ven00] H. Ventsel. *Théorie des probabilités : mathématiques*. MIR Moscou, 2000.

[VM00] T.S. Verma and T.H.Y Meng. Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, 24(2) :47–59, 2000.

[VRB00] E. Veneau, R. Ronfard, and P. Bouthemy. From video shot clustering to sequence segmentation. In *Proc. of ICPR'2000*, pages 254–257, Barcelone, 2000.

[VRSB99] S. Venugopal, K.R. Ramakrishan, S.H. Srinivas, and N. Balakrishan. Audio scene analysis and scene change detection in the mpeg compressed domain. In *Proc. of IEEE 3rd Workshop on Multimedia Signal Processing*, pages 191–196, Copenhagen, Denmark, 1999.

[WBW96] E. Wold, T. Blum, and J. Wheaton. Content-based classification, search and retrieval of audio. *IEEE Multimedia*, 3(3) :27–36, 1996.

[WIB+03] T. Westerveld, T. I. Ianeva, L. Boldareva, A. P. de Vries, and D. Hiemstra. Combining information sources for video retrieval. In *Proc. of TRECVID Workshop'03*, 2003.

[WR00] J. C. Wojdel and L. J. M. Rothkrantz. Silence detection and vowel/consonant discrimination in video sequences. In *Proc. of 8th ICSST'00*, pages 104–109, Australia, 2000.

[WTH00] Y. Wu, Q. Tian, and T.S. Huang. Discriminant algorithm with application to image retrieval. In *Proc. of IEEE CCVPR'00*, 2000.

[YL95] M.M. Yeung and B. Liu. Efficient matching and clustering of video shots. In *Proc. of ICIP'95*, pages 338–341, Washington, 1995.

[Yos96] W.A. Yost. Pitch of iterated rippled noise, July 1996.

[YYL96] M.M. Yeung, B.-L. Yeo, and B. Liu. Extracting story units from long programs for video browsing and navigation. In *Proc. of ICMCS'96*, pages 296–305, Hiroshima, 1996.

[ZF99] E. Zwicker and H. Fastl. Psychoacoustics : facts and models. 1999.

[ZK93] H.-J. Zhang and A. Kankanhalli. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1) :10–28, 1993.

[ZK99] T. Zhang and C.-C. Jay Kuo. Hierarchical classification of audio data for archiving and retrieving. In *IEEE Intl. Conf. on ASSP*, volume 6, pages 3001–3004, 1999.

[ZK01] T. Zhang and C.-C. Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing*, 9(5) :441–457, 2001.

[ZP04] A. Zils and F. Pachet. Automatic extraction of music descriptors from acoustic signals. In *Proc. of ISMIR 04*, 2004.

[ZSN03] A. Zolnay, R. Schulter, and H. Ney. Extraction methods of voicing feature for robust speech recognition. In *Proc. of EuroSpeech'03*, volume 1, pages 497–500, September 2003.

Liste des publications soumises lors de cette thèse

[1] Louis N. and Hanna P., Statistical classification of natural noisy sounds, CBMI 2005, 2005.

[2] Hanna P., Louis N., Desainte-Catherine M. and Benois-Pineau J., Audio features for noisy sounds segmentation, ISMIR 2004, 2004, 1, 120-124

[3] Nesvadba J., Louis N., Benois-Pineau J., Desainte-Catherine M. and Middelink M.K, Low-level cross-media statistical approach for semantic partitioning of audio-visual content in a home multimedia environment, IWSSIP 2004, IEEE, 2004, 1, 235-238

[4] Louis N., Cross-media approach for semantic partitioning of digital multimedia content, Technical Internal Philips Report, 2002.

[5] Benois-Pineau J., Desainte-Catherine M., Louis N., A method for extraction of audio-visual Leitmotif in movies by cross media analysis, EUSIPCO 2002, 2002, III, 345-348