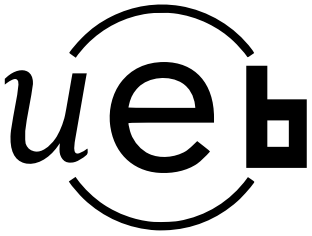
ANNÉE 2013



THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Informatique

École doctorale Matisse

présentée par

Cédric PENET

préparée au centre INRIA Rennes - Bretagne Atlantique
Institut National de Recherche en Informatique et Automatique
Composante Universitaire : IFSIC

De l'indexation d'évènements dans des films – Application à la détection de violence

Thèse soutenue à Rennes
le 10 Octobre 2013

devant le jury composé de :

Laurent MICLET
Professeur émérite à l'université de Rennes 1 / *Président*

Gaël RICHARD
Professeur à TELECOM ParisTech / *Rapporteur*

Régine ANDRE-OBRECHT
Professeur à l'université Paul Sabatier / *Rapporteur*

Benoit HUET
Maitre de conférence à Eurecom / *Examinateur*

Guillaume GRAVIER
Chargé de recherche CNRS / *Co-encadrant*

Patrick GROS
Directeur de recherche INRIA / *Directeur de thèse*

Claire-Hélène DEMARTY
Chercheuse à Technicolor / *Co-directrice de thèse*

Remerciements

Au delà de l'aventure scientifique, ces trois années ont été pour moi l'occasion de côtoyer de multiples personnes que je me dois aujourd'hui de remercier.

Je voudrais commencer par remercier Laurent Miclet, qui m'a fait l'honneur de présider mon jury de thèse. Je souhaite également remercier Gaël Richard et Régine André-Obrecht pour avoir accepter de rapporter mon travail, ainsi que pour leurs regards critiques sur sa qualité. Je remercie également Benoit Huet pour avoir accepter de juger mon travail.

Ayant effectué une grande partie de ma thèse au sein de Technicolor Cesson-Sévigné, je me dois de remercier toutes les personnes avec qui j'ai pu échanger, que ce soit au cours de mon travail, des pauses, des débraillages, des repas, ... Pour ne citer qu'eux, je remercie Philippe, Sandra, Nelly, Cathy, Anne, Izabela, Alberto, Marie et James qui ont partagés nombres de mes repas, Gwen, Franck, Vincent, Matthieu, Julien, Fabrice, Christophe, Jean-Claude, Christel, Ingrid, Lionel, Alexey, Joaquin, Valérie, ... Je n'oublie pas non plus les thésards que j'ai pu croiser, en particulier Alasdair, que j'ai rencontré il y a trois ans et demi le premier jour de notre stage de fin d'étude. Alasdair, je te souhaite bon courage pour la fin de ta thèse, et merci pour ta bonne humeur en toute occasion !

Je remercie aussi les membres de l'équipe INRIA TexMex, au sein de laquelle j'ai effectué l'autre partie de ma thèse, pour leurs remarques, leur bonne humeur lors des séminaires au vert et surtout, pour leur écoute et leur compréhension lors de mes appels à l'aide annuels. Merci tout particulièrement à Camille, pour son amitié et ses discussions lors de mes séjours dans l'équipe.

Il me faut remercier mes trois directeurs de thèse sans qui ce travail n'aurait pas vu le jour. Je remercie donc Guillaume pour sa direction scientifique toujours intéressante et exigeante et le temps qu'il a passé à m'écouter exposer, avec plus ou moins de brio et de clarté, mes nombreuses expériences et idées, pour ses relectures/réécritures d'articles. Je remercie aussi Patrick, qui m'a permis de m'asseoir dans les moments difficiles pour me poser les bonnes questions scientifiques grâce à son "œil extérieur" au problème. Enfin, je me dois de remercier Claire-Hélène pour son suivi et son soutiens, pour les discussions quasi-quotidiennes toujours intéressantes (surtout sur le degrés de violence de telle ou telle scène), pour son soutien, et enfin pour m'avoir appris le travail rigoureux et m'avoir donné l'envie (obligé) d'explorer toutes les pistes d'un problème avant de conclure. Ce travail est aussi un peu le tien, et sans toi, je ne suis pas sûr que sa qualité aurait été la même.

Cette thèse à aussi été pour moi l'occasion de me rendre compte de la valeur de l'amitié. Je remercie donc tous mes amis, qui m'ont à un moment ou à un autre, aidé par leur conversations, les *happy hours*, les soirées, les vacances, le sport, ... aidé à supporter ces trois années intensives. Merci à R& S, Jambon et MDPP, Beun et Momo (mention spéciale pour la correction des fautes d'orthographe !), PA, Seb, Nico, Rico, Loeiz et Marine, Martin, la famille Balbuck (Maxi-White, Marie et la future Nabila), ...

Pour finir, *the last but not the least*, je remercie ma famille, mes parents, mes frères

et surtout Marion, qui me supporte (tout court) depuis maintenant plus de 6 ans !

Table des matières

Introduction

Contexte général de la thèse

Les contenus multimédias sont devenus l'un des principaux moyens de communication dans le monde. La démocratisation des moyens de création a entraîné l'explosion du nombre de contenus multimédias disponibles, ainsi que le développement de nouveaux types de contenus par la même occasion. A titre d'exemple, l'Institut national de l'audiovisuel (INA), chargé en France de la collecte des flux télévisuels et radiophoniques des chaînes et radios au titre du dépôt légal, a vu le nombre de chaînes et radios soumises au dépôt légal monter à environ 120 en 2011 avec l'arrivée des nouvelles chaînes de la télévision numérique terrestre (TNT), ce qui correspond à environ 930 000 heures de contenu sauvegardé tous les ans. La production de films a aussi considérablement augmenté. D'après le Conseil national du cinéma et de l'image animée (CNC), le nombre de longs métrages produits a été multiplié par trois au sein de l'Union Européenne entre 1991 et 2010, passant de 471 à 1 218 longs métrages. Enfin, on peut aussi noter l'apparition de nouveaux modes de consommation tels que la vidéo à la demande (VOD), ou les sites de partage de vidéos amateurs et professionnelles, comme la plateforme YouTube, qui annonce plus de 100 heures de vidéo mises en ligne chaque minute en 2013, ayant ainsi contribué à l'explosion de la quantité de données disponibles.

La démocratisation de la création et de l'accès aux contenus, notamment par internet, a aussi augmenté d'autant la quantité de contenus pouvant heurter la sensibilité d'un jeune public. Cela a, par conséquent, renforcé le besoin en détection des contenus sensibles ainsi que d'une classification des contenus claire et adaptée à la protection des mineurs. Pour le cinéma et la télévision, il existe des échelles de valeurs permettant de classer les films selon leur contenu. Ces échelles ont pour objectif d'indiquer à partir de quelle catégorie d'âge un contenu peut être considéré comme adapté. Ainsi, aux États-Unis, la Motion picture association of America (MPAA) classe les films selon cinq catégories, en fonction de la présence de sexe, de violence, des paroles, ... En France, c'est le Conseil supérieur de l'audiovisuel (CSA) pour la télévision, la VOD, ... et le CNC pour le cinéma qui sont chargés de ce travail. Le dénominateur commun de ces différentes approches est que le très jeune public ne semble pas particulièrement visé par les catégories d'âge. En effet, ces différentes échelles commencent à 10 ans pour le CSA, 12 ans pour le CNC et 13 ans pour la MPAA. Or certains films dits "tout public" ne sont pas adaptés à tous les enfants âgés de 0 à 10 ou 12 ans. Un enfant de 5

ans pourrait donc regarder le film *"Independence Day"* sans problème, bien que celui-ci contienne de nombreuses scènes de destructions massives et de violence. Les échelles de valeurs sont de plus très variables d'un pays à l'autre. Par exemple, la France note les films en mettant l'accent sur la présence de violence, tandis que les États Unis mettent plus l'accent sur la présence de sexe. Cette variabilité des échelles de valeurs est due à une grande subjectivité dans certaines notions rentrant en ligne de compte dans la classification des films, comme la notion de violence, très variable d'une personne à l'autre, et d'une culture à l'autre.

Ainsi, l'augmentation du nombre de contenus multimédias, la diversification des moyens de diffusion de ces contenus, ou le manque de précision dans la classification des films ont accru l'intérêt de développer des techniques d'indexation automatique dédiés à la détection de notions subjectives telles que la violence. Cela permettrait d'accélérer le processus de classification, de proposer de nouvelles applications permettant de mieux contrôler l'accès aux contenus sensibles par des mineurs, ou de nouveaux modes de recherche de contenus dans des bases de contenus. Dans le cadre des films tout particulièrement, la difficulté de la tâche est amplifiée par la grande complexité du contenu, complexité due au travail d'édition effectué sur les films. De plus, cette complexité du contenu, ainsi que la longueur des films, rendent la collecte de données utilisables à des fins de recherche scientifique particulièrement fastidieuse. Ainsi, le défi que nous avons choisi de relever dans cette thèse porte sur la détection automatique d'évènements en général dans les films, et, dans un but de protection du jeune public, nous focalisons nos travaux sur la détection de violence.

Problématiques abordées

Ce cadre applicatif nous amène de nombreuses problématiques scientifiques que nous décrivons dans cette section. Les approches liées à l'indexation de contenus, quel que soit le type de contenus, utilisent en grande majorité des techniques de modélisation fondées sur l'apprentissage automatique. De telles méthodes ont l'avantage de s'appuyer sur des données réelles pour construire un modèle des évènements considérés. Le désavantage est qu'il faut des données pour apprendre le modèle, et que, pour comparer correctement les différentes techniques entre elles, il faut que les modèles soient évalués sur les mêmes données. Dans le cadre des films, et plus particulièrement de la violence dans les films, il n'existe pas de jeu de données public permettant de comparer simplement nos résultats à ceux obtenus par d'autres équipes sur le même sujet. Notre première contribution consiste donc à développer un jeu de données composé de films annotés en terme de concepts audio et vidéo ainsi qu'en terme de violence, et à utiliser ce jeu de données dans le cadre d'une tâche d'évaluation des systèmes de détection de violence que nous proposons à la communauté scientifique dans le contexte de la campagne d'évaluation MediaEval. Cette tâche nous permet ainsi de proposer un lieu d'échange d'idées pour les équipes de recherche travaillant sur le domaine de la détection de violence, et de comparer nos algorithmes aux travaux de ces équipes.

Les films sont un contenu multimodal, c'est-à-dire qu'ils sont composés de plusieurs

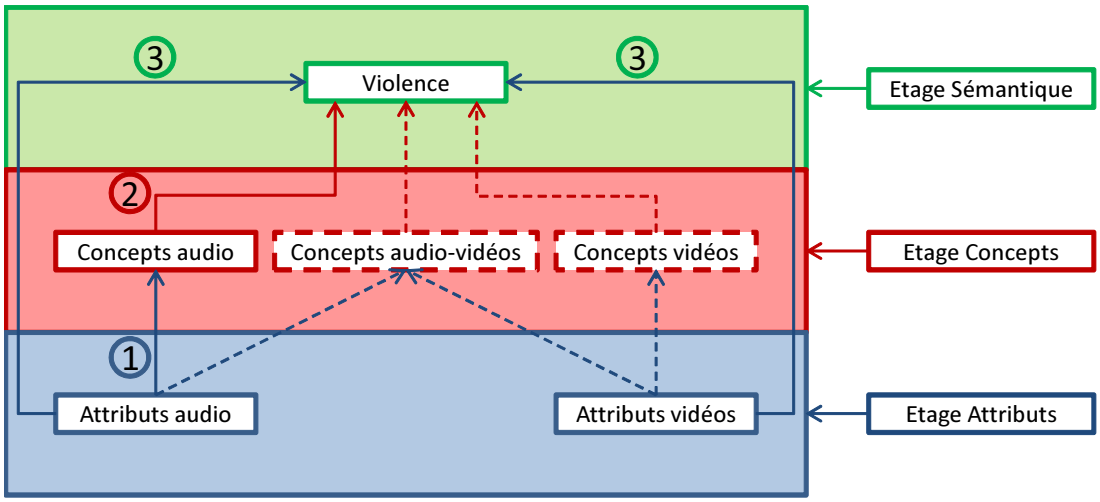


FIGURE 1 – Détection de violence : schéma des problématiques abordées dans ce mémoire. Les flèches pleines correspondent aux approches que nous avons explorées, tandis que les flèches en pointillés correspondent à d'autres approches existantes non étudiées ici.

sources d'information, comme l'audio, la vidéo et les sous-titres, chaque source d'information apportant son lot d'informations, de difficultés, ... Par exemple, il est possible d'extraire des segments de parole de la bande son, ou des images d'armes à feu de la vidéo. Il est aussi possible de détecter la présence de combats en utilisant le son, la vidéo ou les deux. Une partie de la difficulté consiste donc à déterminer quel type d'information on souhaite obtenir, et comment on souhaite l'obtenir, c'est-à-dire à partir de quelle modalité. La figure 1 présente un résumé des différentes approches possibles existantes dans la littérature lorsqu'on se restreint à l'utilisation des modalités audio et vidéo. Les différentes approches abordées dans ce mémoire pour détecter la violence dans les films sont mises en évidence sous forme de flèches pleines et représentent un sous-ensemble de l'ensemble des différentes approches possibles.

Dans un premier temps, nous nous sommes intéressés à la détection d'évènements (ou concepts) audio dans les films (flèche ① sur la figure 1). Si la détection d'évènements audio a beaucoup été étudiée dans la littérature, leur détection dans le cas particulier des films a fait l'objet de beaucoup moins d'études. Cela peut s'expliquer par la difficulté d'obtenir des données utilisables à des fins de recherche, et surtout conséquentes en termes de quantité de données, pour tester les algorithmes. Les évènements intéressants peuvent aussi être très rares, ce qui veut dire que même si l'on dispose de beaucoup de films, il n'y a pas de garantie que les évènements qui nous intéressent soient présents en grande quantité. Nous montrons, dans un premier temps, que les articles ayant trait à la détection d'évènements dans les films font état d'un problème de généralisation spécifique aux bandes son des films pouvant être dû à la complexité des bandes sonores, à la rareté des évènements et à la grande variabilité entre les films. Les différentes approches que nous abordons par la suite sont dédiées à la résolution de ce problème de

généralisation. Nous abordons le problème selon deux axes de recherche. Nous tentons tout d'abord de résoudre le problème en prenant en compte la rareté des évènements, c'est-à-dire en tentant de compenser la rareté des évènements dans les algorithmes utilisés. Notre deuxième approche concerne la réduction de la variabilité entre les films dans les attributs descriptifs du signal audio. Nous faisons ici le pari que la variabilité est transcrite dans les attributs et nous tentons de la réduire en utilisant deux techniques différentes. La première concerne l'utilisation de séquences de mots audio, permettant de représenter des échantillons audio en utilisant la similarité, et la deuxième consiste à directement modéliser cette variabilité sous forme d'un espace vectoriel.

Dans un deuxième temps, nous nous sommes intéressés à la détection de violence (flèches ② et ③ sur la figure 1) et nous explorons deux axes pour détecter la violence : un premier axe monomodal, fondé uniquement sur le canal audio, et un deuxième axe multimodal, fondé sur l'utilisation à la fois de l'audio et de la vidéo. Dans notre première proposition, nous explorons pour commencer l'intérêt d'utiliser un étage intermédiaire de détection de concepts pour détecter la violence à l'aide uniquement de la modalité audio (flèche ②). Cela nous permet d'utiliser les travaux que nous développons dans le cadre de la détection d'évènements audio (flèche ①) à notre avantage. Nous développons ainsi deux approches monomodales différentes. La première approche que nous explorons est empruntée à la reconnaissance de documents textuels similaires, combinée à la représentation par séquences de mots audio. Cette approche nous permet donc de considérer les mots audio comme des concepts obtenus de manière non supervisée. Notre deuxième proposition est inspirée du système développé par Schlüter *et al.* [118, 61], basée sur l'utilisation des sorties de multiples détecteurs de concepts, dont les résultats sont faibles, comme attributs d'un détecteur de violence. Nous intégrons cette idée directement dans le détecteur de concepts que nous développons. Pour finir, nous explorons une approche multimodale, basée sur l'utilisation d'attributs audio et vidéo pour détecter la violence (flèches ③). Nous étudions les méthodes de fusions précoce et tardive existantes dans le cadre de la détection de violence.

Enfin, tout au long de ce mémoire, nous étudions les réseaux bayésiens. Les réseaux bayésiens sont une méthode graphique de représentation de distributions de probabilité, et nous explorons leurs performances dans un contexte de classification. Nous analysons diverses structures de réseaux bayésiens, certaines étant apprises automatiquement, d'autres "faites à la main". En particulier, nous présentons une méthode simple permettant de prendre en compte le contexte des données d'entrées, et nous l'appliquons aux réseaux bayésiens. Nous baptisons les réseaux ainsi créés réseaux bayésiens contextuels, et nous étudions en particulier l'apprentissage de la structure de ce type de réseau.

Organisation du mémoire

Cette thèse s'organise en trois parties. La première partie, composée d'un seul chapitre, est destinée à présenter un état de l'art dédié à l'indexation de contenus. Nous y présentons dans les grandes lignes les systèmes d'indexation, en décrivant les différentes

techniques, et nous présentons en particulier les travaux ayant trait à la détection de la violence. La deuxième partie, dédiée à nos travaux pour la détection d'évènements sonores dans les films, est composée de trois chapitres. Dans le premier chapitre, nous présentons la difficulté de travailler avec des films et nous mettons en évidence le problème de généralisation lié à la grande variabilité entre les bandes sonores des films. Puis, dans le deuxième chapitre, nous décrivons les différentes pistes que nous avons testées pour tenter de résoudre le problème de généralisation. Le troisième chapitre est un bilan de cette partie. La troisième partie, dédiée à la détection de violence, est aussi composée de trois chapitres. Le premier présente la campagne d'évaluation des méthodes de détection de violence dans les films que nous avons mise en place. Le deuxième chapitre présente les différentes approches que nous avons testées sur le sujet. Enfin, le troisième chapitre est un bilan de cette partie. Nous terminons cette thèse en présentant les perspectives ouvertes par nos travaux.

Première partie

État de l'art

Chapitre 1

Indexation automatique de contenus vidéos

L'indexation de contenus multimédias consiste à poser des étiquettes sur des documents multimédias, de façon à pouvoir les trier et les retrouver plus simplement par la suite. L'explosion de la quantité de contenus disponibles a rendu l'automatisation de l'indexation indispensable dans de nombreux domaines applicatifs. On peut par exemple l'employer dans le cadre de la segmentation de matchs de tennis [75, 42], pour la détection d'actions dans des matchs de football [5], ou dans les systèmes de surveillance [32, 35, 33, 26, 100], pour l'analyse de morceaux de musique [114, 111, 112], ou la reconnaissance de genres musicaux [86, 113], de genres vidéos [45, 115], pour la détection d'évènements dans des vidéos [18, 4, 134, 122, 19, 60, 57], ... Les types de contenus sont aussi très variables, allant du simple morceau de musique aux films, en passant par les contenus générés par les utilisateurs (vidéos YouTube par exemple), les dépêches d'informations textuelles, les journaux ou les jeux télévisés, les publicités, ...

Dans ce mémoire, nous nous intéressons en particulier à la détection d'évènements dans les films, avec application à la détection d'évènements audio, et à la détection de violence. Nous proposons donc dans ce chapitre de commencer par décrire ce qu'est un évènement, puis de définir le processus classique de détection d'évènements dans des vidéos avant de présenter un état de l'art sur la détection de violence.

1.1 Sur les évènements dans les contenus multimédias

Il existe, dans la littérature, plusieurs modèles de description des évènements. Par exemple, Xie *et al.* [134] proposent un modèle inspiré des modèles définis par Vendrig et Worring en 2002 [128] et par Westermann et Jain en 2007 [132]. Ils décrivent les évènements grâce à une série de questions appelée 5W1H [1]. Pour présenter leur système, prenons l'exemple *"Une équipe de rugby parisienne remporte à Paris la finale du TOP 14 2013, avec 30 points d'avance sur les challengers toulousains"* pour illustrer les réponses à chacune des questions :

1. De l'anglais : *who ?, when ?, where ?, what ?, why ? and how ?.*

Qui ? Cette question indique le(s) sujet(s) de l'évènement. Dans notre exemple, cela correspond à une équipe de rugby parisienne et leur challengers toulousains.

Quand ? Cette question correspond à l'aspect temporel de l'évènement, c'est-à-dire son caractère passé, présent ou futur, sa fréquence d'apparition, ... Notre équipe de rugby parisienne a gagné en 2013.

Où ? Cette question indique le lieu de l'évènement. Notre évènement exemple se situe à Paris.

Quoi ? Cela correspond à l'évènement en lui-même, comme gagner la finale du TOP 14.

Pourquoi ? Cette question a trait à l'aspect contextuel de l'évènement, plus précisément à la raison pour laquelle l'évènement a lieu. On peut imaginer que notre équipe a gagné parce qu'ils jouaient mieux.

Comment ? Cette question représente l'aspect dynamique de l'évènement, ce qui peut correspondre aux questions "*Comment* était l'évènement ?", ou "*Comment* est-il arrivé ?". La première question correspond à une impression *a posteriori*, tandis que la deuxième correspond à la succession d'évènements ou d'actions ayant mené à l'évènement considéré. Dans le cadre de notre évènement exemple, la réponse à la première question est qu'ils ont gagné avec 30 points d'avance. La première question peut aussi amener un jugement de valeur sur l'évènement, par exemple sur la qualité du jeu développé par les équipes pendant le match. La réponse à la deuxième est qu'ils ont gagné suffisamment de matchs pour arriver en finale ou que la succession d'évènements pendant le match a fait qu'ils ont gagné.

Cette représentation a plusieurs avantages, notamment celui de pouvoir décrire tous les types d'évènements, quels que soient les types de contenus. Elle permet aussi de montrer quelles sont les questions importantes à prendre en compte pour détecter un évènement particulier. Prenons par exemple le cas de la détection de coups de feu dans les films, en nous intéressant particulièrement au coups de feu dans le signal audio, et utilisons la méthode 5W1H pour le décrire :

Qui ? On ne sait pas et cela importe peu en général. En effet, il y a peu de différence entre un coup de feu tiré par une personne ou une autre, en terme de signal.

Quand ? C'est ce que l'on cherche. Détecter les coups de feu revient à détecter les instants temporels où ils se sont produits.

Où ? De même que la première question, on ne sait pas. Le lieu où le coup de feu a été tiré peut avoir son importance sur le son produit, par un exemple un coup de feu tiré en plein air sera différent d'un coup de feu tiré en espace fermé. De même la probabilité d'entendre coup de feu sur un champs de bataille est certainement plus élevée que dans un bureau. Si ce type information est accessible, elle peut être utilisée comme *a priori*, mais ce n'est souvent pas le cas.

Quoi ? Un coup de feu dans le signal audio.

Pourquoi ? C'est en général dans le but de tuer une autre personne. Mais cela peut être aussi une scène montrant des personnes s'entrainant à tirer.

Comment ? Cela dépend du type de film, film de guerre, de gangster, accident, ... La séquence d'actions amenant au coup de feu sera différente selon le type de film. Cela dépend aussi du type d'arme utilisée. Un coup de feu tiré par une arme automatique sera différent de celui tiré par un pistolet.

Ainsi, la technique utilisée par Xie *et al.* [134] nous permet de montrer que, pour les coups de feu et les explosions dans les films, les questions les plus importantes sont *Quand ?*, qui est en partie indépendante de l'évènement considéré, et inhérente à tout problème de détection, *Comment ?* et surtout *Quoi ?*. Les autres questions peuvent aussi être utilisées dans le but d'apporter des informations complémentaires, si celles-ci peuvent être disponibles.

En plus de cette représentation, nous pensons que d'autres attributs additionnels sont importants à prendre en compte. Ainsi, les attributs objectif/subjectif, rare/fréquent, régulier/aléatoire et concept/sémantique peuvent apporter des informations complémentaires sur les évènements. Nous les décrivons dans la suite de cette section en donnant des exemples d'évènements correspondant extraits de la littérature.

1.1.1 Objectifs / subjectifs

La différence entre un évènement objectif et un évènement subjectif tient principalement dans la notion d'interprétation de ce qui peut entrer dans la définition de ces évènements. Un évènement objectif est typiquement un évènement bien défini, qui n'est pas sujet à interprétation par la ou les personne(s) humaine(s) qui pourrai(en)t être amenée(s) à caractériser l'évènement, au contraire des évènements subjectifs. L'avantage des évènements objectifs est de ne pas nécessiter plusieurs avis pour être caractérisés. Ces évènements sont bien souvent l'objet de recherche dans la littérature. Par exemple, Haering *et al.* [59] ont développé un système pour détecter des animaux chassant d'autres animaux dans des documentaires animaliers. Clavel *et al.* [32] se sont quant à eux intéressés à la détection de coups de feu dans des flux audio. En général, les systèmes développés ne sont pas des systèmes dédiés à un seul évènement objectif mais à plusieurs. Ainsi, le système développé par Atrey *et al.* [4] s'intéresse à détection de *discussions, pleurs, coups sur une porte, bruits de pas, marche et course*, celui de Cristani *et al.* [35] à la détection de personnes *recevant ou faisant un appel téléphonique, arrivant premier ou dernier au travail, ou n'arrivant ni premier ni dernier au travail*. Sun *et al.* [122] ont quant à eux utilisé deux bases de données d'évènements tels quel *sortir d'une voiture, serrer une main, ...* Trancoso *et al.* [124, 123, 104, 15, 110] ont utilisé une base d'effets sonores de plus de 47 évènements. Les évènements objectifs sont donc très nombreux, et très variés. Ce sont probablement les évènements les plus présents, car ce sont la majorité des évènements de tous les jours. La majorité des sons peuvent ainsi être apparentés à des évènements objectifs.

L'étude des évènements subjectifs est en revanche plus récente, et donc moins présente dans la littérature. On peut par exemple considérer le champ de la détection d'émotions dans des vidéos [33, 120], les émotions étant très compliquées à définir précisément, ou encore la détection d'actions dans le football [5], ou dans les films [24, 23]. La catégorisation de scènes dans des vidéos peut aussi être très subjective, et dépen-

dante des catégories utilisées : si ce sont des catégories descriptives, telles que *lever de soleil, ville*, la tâche reste objective, mais si l'on souhaite caractériser la violence [106], ou l'esthétique d'une photographie [38], cela devient beaucoup plus subjectif. L'interprétation d'un évènement objectif peut aussi être subjective. Si l'on reprend le cas d'usage de Haering *et al.* [59], un animal en chassant un autre, c'est-à-dire la réponse à la question *quoi ?*, est un évènement objectif. En revanche, la question de savoir *pourquoi ?* cet animal en chasse un autre est subjective. En effet, chasse-t-il pour nourrir ses petits, pour leur apprendre à chasser, pour se nourrir, ... Il est donc important de bien définir ce que l'on souhaite détecter.

1.1.2 Rares / fréquents

La question de la rareté ou de la fréquence des évènements est directement liée au nombre d'occurrences de ces derniers dans le type de contenus, et est directement dépendante de l'application pour laquelle le système est défini : un évènement particulier peut être rare dans une certaine application, c'est-à-dire peu ou pas présent dans les contenus considérés, et fréquent dans une autre, c'est-à-dire apparaissant souvent dans les contenus considérés. La frontière entre un évènement rare et un évènement fréquent est très floue, et il semble difficile de déterminer un seuil à partir duquel on peut considérer qu'un évènement est rare ou fréquent. Un évènement rare peut aussi être vu comme un évènement inhabituel dans le contenu considéré. Un même document peut aussi contenir à la fois des évènements fréquents et des évènements rares. Ainsi, un film contient environ 50 % de silence, ce qui en fait un évènement très fréquent, mais contient en moyenne moins de 3 % de coups de feu ou d'explosions (voir section 2.2.2). Un vol à la tire ou une agression dans une vidéo issue d'un système de surveillance vidéo est aussi un évènement rare. Si la caméra filme une rue non piétonne, le fait de voir une voiture passer est un évènement fréquent, tandis que si la caméra est placée dans une rue piétonne, cet évènement devient très rare. Du fait de la rareté ou de la fréquence des évènements, les difficultés rencontrées ne sont pas les mêmes, et par conséquent, les techniques utilisées peuvent être différentes.

1.1.3 Réguliers / aléatoires

Les évènements que nous définissons en tant que réguliers sont les évènements dont la présence peut servir à structurer les documents considérés. La régularité s'exprime donc par le fait que leur présence est prévisible et systématique. Elle peut aussi aider à reconnaître le document. La régularité peut être inter-documents, comme les génériques de films, la présentation des équipes dans des matchs de sport, ou encore le résumé présent au début des journaux télévisés. Elle peut aussi être intra-documents, comme par exemple les plans sur le présentateur entre deux reportages lors des journaux télévisés, ou encore les "jingles" lors de jeux télévisés. Il faut noter que les évènements réguliers ne sont pas forcément des évènements fréquents. En effet, les génériques de films n'apparaissent qu'une à deux fois dans un film (au début et à la fin). Ces éléments réguliers sont souvent présents dans des documents qui paraissent à dates régulières tels que les

jeux ou les journaux télévisés. En revanche, il est important de bien comprendre que tous les évènements structurels ne sont pas forcément réguliers. Si l'on prend le cas d'un but pendant un match de football, il s'agit bien d'un évènement structurel, c'est le but du match. En revanche, il ne s'agit pas d'un évènement régulier, car la présence de but pendant un match n'est ni prévisible, ni systématique.

Nous appelons évènements aléatoires l'ensemble des autres évènements, c'est-à-dire ceux qui ne sont pas réguliers. Par définition, ils peuvent ou non être présent dans le document, ils peuvent survenir à n'importe quel moment, une ou plusieurs fois. Le nombre d'occurrence de ces évènements dans un document varie en général d'un document à l'autre. C'est par exemple le cas des explosions, des coups de feu ou de la violence dans les films, ou encore du serrage de main entre deux personnes dans un système de surveillance vidéo, des actions dans les rencontres sportives. La majorité des évènements sont en général aléatoires.

1.1.4 Concepts / évènements sémantiques

Nous faisons aussi la distinction entre ce que nous appelons concept et évènement sémantique. La distinction que nous faisons entre ces différents évènements tient dans leur niveau d'abstraction et de complexité. Elle est souvent liée à la définition entre les évènements objectifs et subjectifs. Les évènements subjectifs sont souvent complexes à définir et très abstraits, et sont souvent des évènements sémantiques. Les évènements objectifs, quant à eux, sont souvent des concepts. On peut aussi noter le fait que les évènements sémantiques ont souvent une signification importante dans le contenu, et qu'ils sont souvent constitués d'une série de concepts. L'évènement sémantique constitue donc l'interprétation de l'action représentée par la série de concepts. Ainsi, un coup de feu ou une explosion, évènements objectifs, peuvent être considérés comme des concepts dans le cas de films, tandis que la violence sera considérée comme un évènement sémantique. La parole peut-être considérée à la fois comme un évènement sémantique et un concept. Si l'on cherche uniquement à détecter la présence ou non de parole, alors on peut dire que la parole est un concept, mais si l'on cherche à la transcrire, alors la parole devient un évènement sémantique, car constitué d'une série de phonèmes. Ainsi, la classification concept ou évènement sémantique peut dépendre du contexte dans lequel on se situe.

1.2 Processus classique d'indexation dans la vidéo et spécificités liées aux modalités

Il existe un grand nombre de publications ayant trait à la détection d'évènements dans des vidéos, souvent fondées soit sur la vidéo, soit sur l'audio, et plus rarement multimodales. Les systèmes développés dans la littérature ont cependant tous la même base que nous avons tenté de résumer sur la figure 1.1. Ainsi, quel que soit le type d'information utilisé dans la vidéo ou quel que soit l'évènement considéré, le processus est souvent découpé en deux étapes : le contenu passe systématiquement par une phase d'extraction d'attributs, permettant de décrire le contenu, et par une phase de

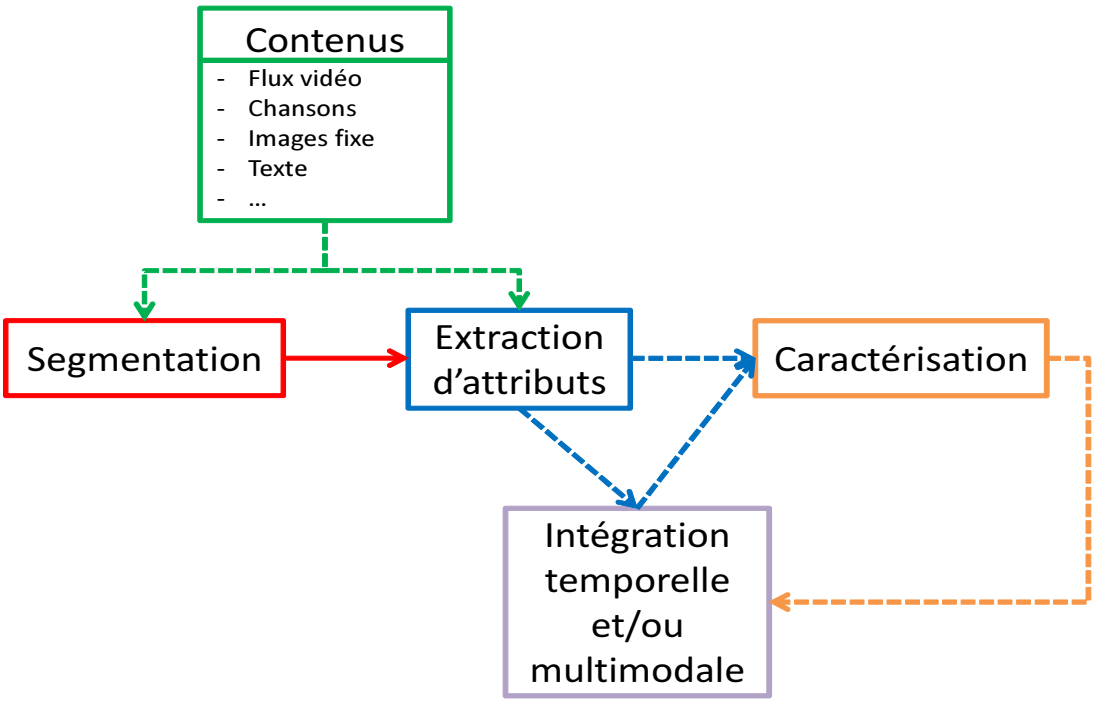


FIGURE 1.1 – Processus classique d'indexation d'évènements. Les flèches en pointillés correspondent à des étapes facultatives (ex : l'étape d'intégration après l'étape de caractérisation), ou aux différentes possibilités en sortie d'une étape (ex : le contenu peut soit être segmenté avant l'extraction d'attributs, soit passer directement dans la phase d'extraction d'attributs). Les flèches pleines correspondent à une relation obligatoire entre deux étapes (ex : si le contenu est segmenté, alors il y a ensuite automatiquement extraction d'attributs).

caractérisation, permettant d'assigner une(des) étiquette(s) au contenu. Ces phases sont en général agrémentées d'une phase de segmentation, permettant de "découper" le contenu en segments, d'une phase d'intégration temporelle, permettant de prendre en compte d'une manière ou d'une autre l'évolution temporelle dans le contenu, et d'une phase d'intégration multimodale, dans le cas où plusieurs modalités sont utilisées. Si les phases de caractérisation, d'intégration temporelle et d'intégration multimodale sont généralement indépendantes de la modalité d'entrée du système, les phases d'extraction d'attributs et de segmentation du flux sont très différentes selon que le système considéré est basé sur l'audio ou la vidéo. Nous proposons donc dans cette section de présenter brièvement les spécificités de l'audio et de la vidéo concernant la segmentation du flux et l'extraction d'attributs, puis de présenter les techniques d'intégration temporelle et de caractérisation du contenu courantes avant de présenter les techniques d'intégration multimodale.

1.2.1 Segmentation du flux et extraction d'attributs

Les étapes de segmentation du flux et d'extraction d'attributs sont de fait indissociables lorsqu'elles sont toutes les deux présentes dans le système. En effet, les attributs extraits dépendent en général de la segmentation effectuée en amont, dans la mesure où ils seront extraits de chaque segment obtenu lors de l'étape de segmentation. Il est important que ces attributs portent l'information nécessaire à la détection des évènements que l'on considère, puisqu'ils seront ensuite utilisés par l'étape de caractérisation. Nous détaillons dans la suite les différents types d'attributs en fonction des types de segmentation du flux, en nous attardant par endroits sur ceux que nous utilisons dans ce mémoire.

1.2.1.1 Signal vidéo

Le signal vidéo est une suite d'images pouvant être considérées soit indépendamment les unes des autres, soit par groupe d'images, c'est-à-dire par deux ou trois, par plans, par scènes, ... Chaque niveau de segmentation permet d'extraire différents types d'attributs. Dans la suite, nous présentons les grandes familles d'attributs qui peuvent être extraits aux différents niveaux de segmentation, en illustrant notre propos, quand c'est possible, par des exemples extraits des attributs que nous utilisons dans ce mémoire, ou par des exemples courants de la littérature.

Image : Les images du flux vidéo contiennent en elles-mêmes beaucoup d'information. Citons pour commencer les attributs liés à la couleur. Il est effet possible d'extraire des statistiques ou des histogrammes sur les couleurs présentes dans l'image. On peut aussi extraire d'autres informations telles que la proportion de pixels de couleur sang [99, 106] ou encore des attributs liés à l'harmonie couleur [6]. De la même façon, des attributs analysant les différentes textures dans les images peuvent être extraits [131]. Il est aussi possible de d'effectuer une segmentation de l'image, en utilisant par exemple les couleurs ou les gradients de l'image. Cela peut par exemple être utilisé pour extraire des régions d'intérêt (ROI) dans l'image. Enfin, les attributs images les plus populaires ces dernières années restent peut-être ceux basés sur la transformée SIFT [2] [91]. Les descripteurs SIFT sont des descripteurs invariant aux changements d'échelles extraits autour de points d'intérêt dans l'image. Si ces attributs décrivent assez efficacement les images fixes, ils ne prennent en revanche pas en compte l'évolution temporelle des images.

Groupe de N images : Pour prendre en compte l'évolution temporelle des images, on peut extraire des attributs sur des groupes de N images consécutives. En général, le rythme d'extraction des attributs est d'un attribut par image, mais N images sont nécessaires pour les extraire. Le principal attribut utilisé dans ce cas est l'activité entre deux images, correspondant plus ou moins à la quantité de mouvement qu'il y a entre deux images. L'activité entre deux images correspond en général à la norme moyenne

2. De l'anglais *Space-Invariant Feature Transform.*

des vecteurs mouvements issus du flot optique entre les images, comme par exemple dans [24, 106]. Il est aussi possible d'utiliser le champ de vecteurs mouvements comme attribut directement. On peut aussi citer l'extension de la transformée SIFT au domaine spatio-temporel introduite par Laptev [83], permettant de détecter des points d'intérêt à la fois dans l'espace et dans le temps. Cette transformée a été utilisée notamment dans [40]. Comme exemple supplémentaire, on peut aussi citer la détection de grandes variation de luminance entre trois images successives [99, 108, 106].

Plans vidéos ou scènes : Une vidéo est en général constituée d'une série de plans successifs. De nombreuses techniques de segmentation en plans ont été développées, parmi lesquelles celles qui comparent les histogrammes couleur d'images successives. La durée de ces plans peut ainsi servir d'attribut dans un système. Des plans courts peuvent ainsi indiquer que la vidéo contient de l'action [24]. On peut aussi caractériser les plans par ce qu'il y a dans les images : on peut parler de gros plans, de plans larges, ... Les plans vidéos peuvent aussi être groupés par scènes en analysant par exemple la cohérence couleur entre les différents plans [102]. On peut alors extraire des attributs temporels de la séquence de plans successifs constituant la scène [24].

1.2.1.2 Signal audio

Le signal audio est un signal unidimensionnel non stationnaire variant très rapidement. Il est donc compliqué d'utiliser les techniques classiques de traitement du signal faisant l'hypothèse de stationnarité du signal, telles que la transformée de Fourier directement sur le signal entier. Ainsi, pour permettre d'extraire des attributs descriptifs du signal, ce dernier est découpé en fenêtres d'une durée variant entre en général 10 et 50 ms en fonction des systèmes, avec un taux recouvrement entre deux fenêtres audio successives variant de 30 % à 50 %. L'hypothèse de stationnarité du signal est donc faite sur ces petites fenêtres. Il existe cependant certains attributs pouvant se calculer sur des fenêtres d'analyse plus longues, de l'ordre d'une seconde, notamment les attributs dits temporels. Concernant les attributs audio, Essid [49] présente de manière relativement exhaustive les principaux attributs utilisés dans la littérature, qu'il découpe en quatre catégories : temporels, spectraux, cepstraux et perceptuels. Nous listons ici certains de ces attributs. Pour une liste plus complètes et les formules mathématiques correspondantes, se reporter à [49].

Attributs temporels : Les attributs temporels sont ceux mesurés directement sur le signal audio, sans transformation préalable de celui-ci. On compte parmi eux le taux de passage par zéro (ZCR), indiquant si le signal a une plus ou moins haute fréquence. Cet attribut peut être calculé sur des fenêtres longues ou courtes. Il est également classique d'extraire des attributs liés à la modulation d'amplitude, tels que la fréquence et l'amplitude de l'enveloppe de modulation, calculées sur des fenêtres longues, ou encore les différents moments temporels, pouvant être calculés à court et long termes, sur le signal audio lui-même ou sur son enveloppe de modulation. Nous utilisons très

peu ces attributs dans ce mémoire, mais ils sont utilisés dans le cadre de la segmentation parole/musique [114, 111, 112].

Attributs spectraux : Les attributs spectraux sont calculés à partir de la transformée de Fourier du signal audio sur des fenêtres courtes et sont bien plus utilisés que les attributs temporels. Parmi les plus utilisés, on peut noter l'énergie ou encore les moments spectraux donnant une information sur la forme du spectre dans la fenêtre d'analyse considérée, tels que le barycentre du spectre, donnant une information similaire au taux de passage par zéro, la variance du spectre, donnant une indication sur son étalement autour du barycentre, l'asymétrie, donnant une information de symétrie du spectre par rapport au barycentre. D'autres attributs décrivant la forme spectrale existent par ailleurs. La platitude du spectre, définie comme le ratio entre les moyennes géométrique et arithmétique du spectre, indique si le spectre de la fenêtre considérée ressemble à du bruit blanc (spectre plat). La fréquence de coupure permet aussi de caractériser l'importance des sons de haute fréquence dans la fenêtre. Il existe aussi des attributs dynamiques permettant pour chaque fenêtre d'analyse d'indiquer les variations du spectre (flux spectral) ou le suivi de la fréquence fondamentale (pitch).

Attributs cepstraux : Les attributs cepstraux sont extraits du cepstre du signal, défini comme étant la transformée de Fourier inverse du logarithme du spectre d'amplitude du signal. Cette transformée permet de représenter les variations du spectre d'amplitude et est beaucoup utilisée en reconnaissance de la parole par exemple. Il en existe de nombreuses variantes, mais la plus utilisée est probablement la variante MFCC[3]. Pour calculer les MFCC, le spectre d'amplitude est passé dans un banc de filtres triangulaires uniformes dans l'échelle de Mel, permettant ainsi d'obtenir une représentation assez fine dans les basses fréquences et plus grossière pour les hautes fréquences, à l'image de la perception humaine des sons. Une transformée en cosinus est ensuite appliquée aux énergies ainsi extraites, et les MFCC sont les coefficients de la transformée. Les MFCC sont utilisés dans presque tous les systèmes d'indexation audio, en particulier en reconnaissance de la parole ou du locuteur, ce qui en fait un jeu d'attributs presque incontournable.

Attributs perceptuels : Les attributs perceptuels sont des attributs liés à l'audition humaine. On peut noter parmi eux les attributs loudness et sharpness par exemple. Nous ne les avons pas utilisés dans ce mémoire.

Plus généralement, et pour tous les types d'attributs, les fenêtres d'analyse courtes n'étant pas très pratiques à utiliser, il est aussi possible de segmenter le signal audio différemment, en extrayant les moments d'attaque du signal audio [7] ou en détectant les ruptures de stationnarité dans le signal [2] par exemple. En combinant ces techniques à la détection de silence [69, 107], on peut obtenir une segmentation en unités logiques de tailles variables en plus de la découpe en fenêtres d'analyse courtes. L'intégration

3. De l'anglais *Mel-Frequency Cepstral Coefficients*.

temporelle peut ensuite être utilisée pour intégrer les attributs extraits selon cette nouvelle segmentation automatique. Les attributs extraits du signal audio ou vidéo sont ensuite utilisés par des méthodes de caractérisation du contenu.

1.2.2 Caractérisation du contenu

L'étape de caractérisation du contenu permet d'associer aux différents échantillons composant le contenu multimédia une ou des étiquettes représentant ces échantillons. Cette étape est aussi appelée étape de classification. Les attributs extraits des signaux audio et vidéos sont ici utilisés comme information pour déterminer les étiquettes à associer aux échantillons.

Les techniques de caractérisation du contenu utilisent en grande majorité les techniques d'apprentissage automatique permettant de construire des modèles mathématiques à partir des données disponibles pour un problème particulier. Les données se présentent en général sous la forme d'un ensemble de N échantillons, chaque échantillon s_i étant représenté par un couple $(y_i, \mathbf{X}_i)$, $\mathbf{X}_i = \{x_{i1}, \ldots, x_{ij}, \ldots, x_{iD}\} \in \mathbb{R}^D$ étant un vecteur d'attributs de dimension D extrait du signal correspondant à l'échantillon et y_i étant une ou plusieurs étiquettes (aussi appelées classes) associées à l'échantillon. Les techniques de caractérisation se divisent en deux grandes catégories. La première catégorie est celle des techniques descriptives. Ces techniques produisent des modèles descriptifs des données, c'est-à-dire des modèles cherchant à décrire la distribution des données disponibles. Parmi les techniques descriptives, on peut citer la technique des K plus proches voisins (KNN), les mélanges de gaussiennes (GMM), les modèles de Markov cachés (HMM) ou les réseaux bayésiens (BN) ou les techniques de quantification vectorielle comme la technique des K moyennes (KMeans). La deuxième catégorie est celle des techniques dites discriminantes. Les modèles issus de ces techniques cherchent à séparer les données en fonctions des classes y_i des échantillons, c'est-à-dire trouver une fonction permet de prédire la classe d'échantillons nouveaux. Ainsi, l'idée générale derrière l'apprentissage de ces modèles peut se résumer comme suit : on cherche f, application de $\mathbb{R}^D$ dans $\mathbb{R}^c$, c étant le nombre d'étiquettes possibles, tel que $y_i \simeq f(\mathbf{X}_i)$, $\forall i \in [1, N]$. Une fois apprise, l'application f peut ensuite être utilisée pour associer des étiquettes à des échantillons dont on ne connaît pas l'étiquette. Parmi ces techniques, on compte notamment les arbres de décisions (DT), les réseaux de neurones (NN) ou les machines à vecteurs supports (SVM).

Il est aussi possible de faire la distinction entre une utilisation supervisée ou non supervisée des techniques d'apprentissage. Utiliser une technique de manière non supervisées revient à prendre en compte uniquement la partie $\mathbf{X}_i$ des échantillons, c'est-à-dire que les étiquettes des échantillons ne sont pas nécessaires pour apprendre un modèle. En revanche, une utilisation supervisé signifie qu'il est nécessaire d'utiliser les étiquettes y_i pour apprendre un modèle. Les techniques d'apprentissages automatiques peuvent en général être utilisées soit de manière supervisée, soit de manière non supervisée.

Toutes ces techniques sont couramment utilisées dans l'état de l'art, c'est pourquoi nous proposons dans ce mémoire de nous limiter à présenter brièvement les GMM, les BN et les BN , qui sont les techniques que nous utilisons le plus dans nos travaux.

1.2.2.1 Mélange de gaussiennes

L'idée derrière les GMM est de modéliser la distribution des échantillons par un mélange de distributions de type gaussiennes multivariées :

$$p(\mathbf{X}) = \sum_{i=1}^{M} \pi_i \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1.1}$$

où :

$$\mathcal{N}(\mathbf{X}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{X}-\boldsymbol{\mu}_i)} \tag{1.2}$$

$$\text{et } \sum_{i=1}^{M} \pi_i = 1 \tag{1.3}$$

Ainsi, le GMM décrit la distribution des données dans l'espace, sans nécessité de connaître les étiquettes. Les paramètres π_i, $\boldsymbol{\mu}_i$ et $\boldsymbol{\Sigma}_i$ sont appris à l'aide d'un algorithme d'espérance-maximisation (EM). Les GMM peuvent être manipulés au travers de leur super vecteur de moyennes $\boldsymbol{\mu}$, concaténation des vecteurs moyennes $\boldsymbol{\mu}_i$, et de leur super matrice de variance $\boldsymbol{\Sigma}$, contenant les matrices de variance $\boldsymbol{\Sigma}_i$ sur la diagonale.

Il est possible de se servir des GMM dans un contexte supervisé, en séparant les données d'entrée $\mathbf{X}_i$ selon leurs étiquettes, et en apprenant un GMM par étiquette. Les nouveaux échantillons sont ensuite testés par chacun des modèles, et l'étiquette du modèle donnant la vraisemblance maximale est associée aux échantillons. Ce type de techniques a surtout été utilisé dans les débuts de la classification, majoritairement en audio [116, 95, 32, 33], mais aussi en image, comme dans [51] ou l'avant et l'arrière plans sont modélisés par des GMM. Plus récemment, les GMM ont surtout été utilisés en tant que densités de probabilité, comme dans [69]. Dans ces travaux, les densités de probabilité associées aux états d'un HMM sont modélisées par des GMM à 8 gaussiennes. Dans [74, 73, 72, 71, 93, 94], le GMM sert de base à la modélisation de la variabilité dans des expériences de reconnaissance du locuteur. On peut aussi se servir du GMM comme d'une méthode de partitionnement flou des données comme dans [50] par exemple.

1.2.2.2 Réseaux bayésiens

Il existe dans la littérature de nombreux tutoriels dédiés aux BN, tels que [21], ou encore [97]. On peut aussi citer le livre de Naïm *et al.* [98], en français, comme étant la référence théorique la plus complète sur le sujet à notre connaissance.

A l'instar des GMM, l'idée des BN est de représenter la probabilité jointe $P(\mathbf{X})$ d'un ensemble de variables $\mathbf{X} = \{x_1, \ldots, x_D\}$ en utilisant un graphe acyclique dirigé (DAG) permettant de représenter les dépendances entre les variables et de simplifier le calcul de $P(\mathbf{X})$. Les BN font ainsi partie de la famille des techniques descriptives, car il n'est pas nécessaire d'utiliser les étiquettes pour les utiliser. La théorie des probabilités nous dit que la probabilité jointe de $\mathbf{X}$ s'écrit :

$$P(\mathbf{X}) = p(x_D|x_1, \ldots, x_{D-1}) p(x_{D-1}|x_1, \ldots, x_{D-2}) \ldots p(x_2|x_1) p(x_1)$$

L'utilisation d'un graphe G pour représenter les dépendances entre les variables permet donc de simplifier la formule :

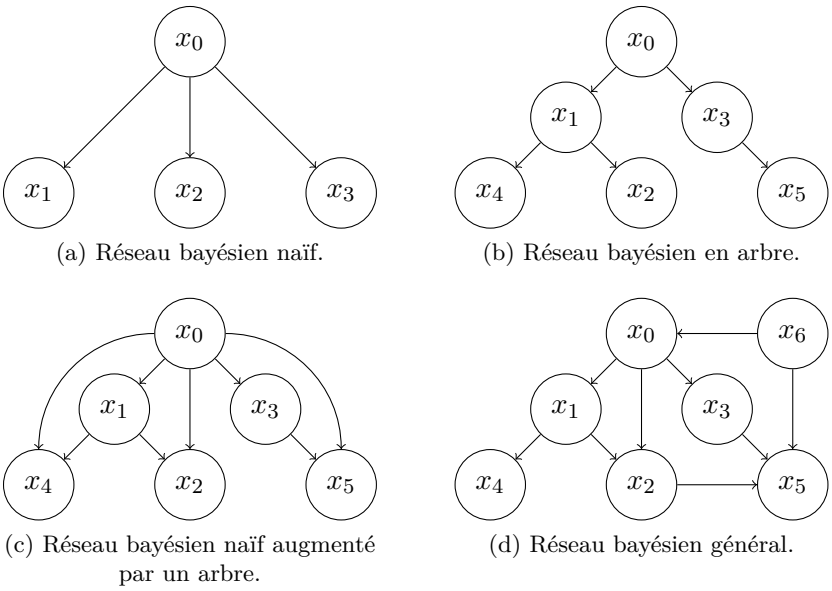
$$P_G(\mathbf{X}) = \prod_{i=1}^{D} p_G(x_i|Pa_G(x_i)) \tag{1.4}$$

où $Pa_G(x_i)$ représente les parents de la variable x_i dans le graphe G. Chaque variable correspond à un nœud du graphe G et les dépendances entre les variables sont représentées par les arcs du graphe. Ainsi, chaque nœud permet de stocker $p_G(x_i|Pa_G(x_i))$, quantité apprise à partir des données d'apprentissage directement. Dans le cas où les variables sont discrètes et complètement observées, $p_G(x_i|Pa_G(x_i))$ correspond à une table de probabilité contenant les fréquences d'apparition de chaque combinaison des états de x_i et des états de $Pa_G(x_i)$ apprises par simple comptage dans la base d'apprentissage. Dans le cas où certaines variables ne sont pas complètement observées, le comptage est estimé par un algorithme d'espérance-maximisation. Dans le cas où les variables sont continues, les nœuds contiennent les paramètres des distributions estimées sur les données.

L'inférence dans les réseaux bayésiens permet de les utiliser dans un contexte de classification. En effet, l'inférence permet, lorsqu'on connaît les paramètres du réseau, d'évaluer $P(x_i|\mathbf{X} \setminus x_i)$. Il est possible, comme dans le cas des GMM, d'utiliser les BN dans un contexte de classification, les étiquettes associées aux échantillons y_i pouvant être vues comme une variable aléatoire au même titre que les attributs x_i. Ainsi, leur densité de probabilité jointe $P(y, \mathbf{X})$ peut être représentée par un BN. La phase d'inférence permet ensuite d'estimer la probabilité de chaque étiquette possible par rapport aux valeurs des attributs de nouveaux échantillons :

$$P(y = y_i|\mathbf{X}) = \frac{P(y = y_i, \mathbf{X})}{\sum_{j=1}^{c} P(y = y_j, \mathbf{X})} \tag{1.5}$$

Les BN sont très dépendants de la structure de leur graphe, qui influence grandement la forme de la distribution représentée. Les figures 1.2a, 1.2b, 1.2c et 1.2d présentent des exemples de structures particulières de réseaux bayésiens. La structure naïve (figure 1.2a) est probablement la forme de structure la plus répandue car elle est simple à mettre en œuvre. Elle suppose une indépendance conditionnelle entre les variables $\{x_i,\ i \neq 0\}$, par rapport à la variable x_0. En dehors de la structure naïve, il peut en revanche être très difficile de créer la structure à la main, surtout en présence d'un grand nombre de variables. C'est pourquoi des algorithmes d'apprentissage de structure ont été développés. Il existe globalement deux familles d'algorithmes : les algorithmes causaux et ceux basés sur l'optimisation d'un score. La première famille est représentée essentiellement par les algorithmes PC, proposé par Spirtes *et al.* [121] et IC*, proposé par Pearl [103]. Le graphe est construit en mesurant l'indépendance entre les variables grâce au test d'indépendance du χ^2, par exemple. Le principal problème de ces méthodes est qu'elles sont très longues à apprendre en présence d'un grand nombre de variables. Les algorithmes fondés sur des scores sont quant à eux beaucoup plus développés. L'idée est de trouver la structure qui minimise ou maximise un score



(a) Réseau bayésien naïf.

(b) Réseau bayésien en arbre.

(c) Réseau bayésien naïf augmenté par un arbre.

(d) Réseau bayésien général.

FIGURE 1.2 – Exemples de structures particulières de réseaux bayésiens.

représentant l'adéquation de la structure aux données pondérée par terme représentant la complexité du graphe. On peut citer par exemple le score BIC :

$$S_{BIC} = \underbrace{\log P_G(\mathbf{X}|\theta_G)}_{\text{Adéquation aux données}} - \underbrace{\frac{1}{2} card(\theta_G) \log N}_{\text{Complexité du graphe}} \tag{1.6}$$

où θ_G représente l'ensemble des paramètres du modèle, et N le nombre total d'échantillons utilisés pour apprendre la structure. Le caractère surexponentiel de l'espace des structures de graphe de type DAG oblige de plus à utiliser des heuristiques permettant de limiter la taille de l'espace à parcourir. On peut notamment citer les méthodes se focalisant sur des arbres (figure 1.2b) telles que la méthode MWST [4] construisant l'arbre de poids minimum à partir des données. Les réseaux naïfs peuvent aussi être augmentés par un arbre (TAN, figure 1.2c) ou une forêt d'arbres (FAN) pour relaxer l'hypothèse d'indépendance entre les variables par rapport aux étiquettes faite par le réseau naïf [92]. Ce type de structures est sensé être plus efficace que le réseau naïf dans un contexte de classification. On peut aussi utiliser des algorithmes moins stricts sur la structure tels que l'algorithme K2 [34], nécessitant un ordonnancement des nœuds, ou l'algorithme GES [29], qui est un algorithme génétique glouton de type *Hill-Climbing*, aboutissant à des structures sans forme particulière (figure 1.2d). Il est enfin important de noter que ces scores sont des scores descriptifs, considérant la variable classe comme étant une variable au même titre que les attributs. Il existe dans la littérature des scores dits discriminants [58, 5], c'est-à-dire optimisant la vraisemblance conditionnelle

4. Pour *Minimum Weight Spanning Tree*.

du nœud classe par rapport aux attributs, mais nous ne les avons pas explorés dans ce mémoire.

Dans la littérature, les réseaux bayésiens sont principalement utilisés pour décrire des données, comme dans le cas de systèmes de diagnostique médical [39]. Dans le domaine de la classification, ils sont surtout utilisés avec des jeux de données très simples tels que dans [9, 27, 28, 58], permettant principalement d'évaluer rapidement sur des données simples les performances des BN par rapport à d'autres techniques. Leur utilisation dans le cadre de la détection d'évènements multimédias est plus rare. On peut tout de même citer [63], mettant en œuvre plusieurs niveaux hiérarchiques de BN pour segmenter les flux télévisuels. Giannakopoulos *et al.* [54, 109, 55, 108] utilisent quant à eux un système hybride KNN/BN pour détecter des évènements audio tels que des coups de feu ou de la musique dans des films. L'influence de l'apprentissage de structure est aussi étudiée dans [5] dans le cadre de la détection d'actions dans le football. Cette étude montre qu'il est possible d'améliorer les performances d'un BN naïf par apprentissage de structure.

Il existe de nombreuses variantes des BN dans la littérature, comme par exemple les réseaux bayésiens dynamiques [52], ou encore les réseaux Multinet [9], consistant à apprendre un réseaux bayésien par classe. On peut aussi citer les réseaux utilisant des variables latentes [10, 48, 46, 137, 47], permettant d'introduire des variables dont ne connaît aucune information, et qui servent à améliorer la représentation des données, ou d'extraire des attributs des concepts inconnus.

1.2.2.3 Machine à vecteurs supports

Les SVM sont probablement les modèles les plus utilisés dans la littérature sur la détection d'évènements et il existe de nombreux articles décrivant les mathématiques associées à cette technique tels que [16, 67, 8, 20]. L'idée de base est assez simple. Si $y_i \in \{-1; +1\}$, il s'agit de trouver l'hyperplan w qui sépare au mieux les données appartenant à la classe $+1$ de celles appartenant à la classe -1, c'est-à-dire l'hyperplan maximisant la distance des échantillons à l'hyperplan de sorte que les échantillons d'une même classe se retrouve du même côté de l'hyperplan. La distance entre l'hyperplan et les échantillons les plus proches de ce dernier est appelée la marge. Dans le cas où les données sont non séparables, on a recours à des fonctions noyaux, permettant de transposer les échantillons dans un espace de dimension potentiellement infinie, dans lequel on cherche l'hyperplan de séparation des échantillons. Si le nombre d'étiquettes est supérieur à deux, c'est-à-dire $c > 2$, on divise le problème en plusieurs sous-problèmes bi-classes, soit une classe contre une classe (stratégie dite *"one versus one"*), soit une classe contre toutes les autres (stratégie dite *"one versus all"*).

Il existe plusieurs formulations mathématiques des SVM. Nous présentons ici très brièvement la formulation la plus utilisée appelée C-SVC [5]. Le problème d'optimisation à résoudre pour trouver l'hyperplan se traduit mathématiquement dans sa forme

5. De l'anglais *C Support Vector Classification.*

primale par :

$$\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{i=1}^{N}\xi_i \tag{1.7}$$
$$y_i\left(\mathbf{w}^t\phi(\mathbf{X}_i) + b\right) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

où $\phi(\cdot)$ correspond à une fonction mapping, les ξ_i sont des variables permettant de relâcher la contrainte sur l'hyperplan et C est un hyperparamètre contrôlant le poids accordé aux variables ξ_i. Cette forme étant difficile à optimiser, on utilise la forme duale de ce problème, définie à l'aide des multiplicateurs de Lagrange $\boldsymbol{\alpha}$:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \tag{1.8}$$
$$\text{sous la contrainte : } \mathbf{y}^T\boldsymbol{\alpha} = 0 \tag{1.9}$$
$$0 \leq \alpha_i \leq C \tag{1.10}$$

où $\mathbf{e}$ est un vecteur composé uniquement de 1, $\mathbf{Q}$ est une matrice telle que $Q_{ij} = y_i y_j K(\mathbf{X}_i, \mathbf{X}_j)$ et $K(\mathbf{X}_i, \mathbf{X}_j) = \phi(\mathbf{X}_i)\phi(\mathbf{X}_j)$ est la fonction noyau appliquée aux données. Les principales fonctions noyaux sont le noyau linéaire $K_L(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{X}_i \cdot \mathbf{X}_j$ et le noyau gaussien $K_{RBF}(\mathbf{X}_i, \mathbf{X}_j) = e^{-\gamma||\mathbf{X}_i - \mathbf{X}_j||^2}$.

De nombreuses variantes des SVM existent dans la littérature, comme par exemple les SVM multinoyaux [122], remplaçant le noyau simple par un mélange de noyaux, ou encore les SVM optimisant le classement des données (Rank-SVM) [68].

Les SVM sont très employés dans la littérature comme technique de classification, à la fois en audio et en vidéo. On peut par exemple citer [85] ou [124, 123, 15, 104, 110] qui utilisent les SVM pour discriminer des effets sonores, [13] qui les utilisent pour classer des films par genre, ou [126] qui cherchent à identifier des locuteurs dans des talk-shows.

1.2.3 Intégration temporelle

L'étape de caractérisation du contenu est souvent, mais pas nécessairement, précédée ou suivie d'un étape d'intégration temporelle permettant de prendre en compte l'information temporelle du contenu dans le système. Joder *et al.* [69] présentent dans leur article les types d'intégration temporelle les plus utilisés en les appliquant à l'audio. Il est important de noter que ce qui est valable pour le signal audio est aussi valable pour le signal vidéo en ce qui concerne l'intégration temporelle. Il existe deux niveaux d'intégration temporelle : l'intégration précoce, généralement effectuée à l'aide des attributs, et l'intégration tardive, effectuée après l'étape de classification. Nous listons ici les principales techniques présentées dans [69].

Intégration précoce :

Cela consiste à agréger les attributs dès l'étape d'extraction de ces derniers sur des fenêtres plus importantes que celles sur lesquelles ils ont été extraits. Par exemple, pour

chaque plan vidéo on peut imaginer extraire un attribut représentant l'activité globale du plan à partir des activités extraites pour chaque image, ou on pourrait extraire un attribut représentant la platitude du spectre toutes les secondes en intégrant les valeurs de platitude spectrale extraites toutes les 20ms. Il existe pour cela plusieurs techniques :

Statistiques - L'intégration par des statistiques est la plus commune. Il s'agit d'extraire des statistiques de l'attribut considéré sur la durée voulue, par exemple un plan vidéo. On extrait généralement la moyenne et/ou la variance de l'attribut.

Modèles auto-régressifs - Cela consiste à remplacer la séquence de valeurs de l'attribut par les coefficients d'un modèle auto-régressif appris sur cette séquence.

Attributs spectraux - Cela consiste à calculer des attributs spectraux sur le spectre de la séquence de valeurs de l'attribut.

Empilement d'attributs - C'est la méthode la plus simple. Elle consiste à concaténer tous les vecteurs d'attributs de la séquence considérée. C'est la méthode qui augmente le plus rapidement la taille du vecteur d'attributs. Elle impose par ailleurs une segmentation fixe du signal.

Intégration tardive :

L'intégration tardive s'utilise une fois la décision prise, c'est-à-dire après l'étape de caractérisation du contenu. Cela permet d'obtenir une deuxième décision sur des fenêtres plus importantes ou de filtrer les décisions prises. Nous détaillons deux façons de faire :

Utilisation de classifieurs - Il est possible d'utiliser la séquence de décision comme attributs d'entrée d'un deuxième classifieur tel que les HMM ou les réseaux bayésiens pour prendre une deuxième décision sur une séquence de décisions d'une taille plus ou moins importante.

Filtrage temporel - L'utilisation de filtres temporels est aussi envisageable. Cela permet soit de prendre une décision sur un horizon temporel plus grand, soit de filtrer les décisions déjà prises en utilisant une fenêtre d'analyse glissante. Les filtres les plus populaires sont le vote majoritaire, permettant de prendre la décision majoritaire dans la fenêtre d'analyse, et le filtre moyenneur, dans le cas où les décisions sont des valeurs continues telles que des probabilités.

Dans leur article, Joder *et al.* [69] comparent l'intégration précoce et tardive. Ils indiquent notamment qu'il n'y a pas beaucoup de différence de résultats entre les différentes techniques, ni entre l'intégration tardive et précoce. En revanche, il semble que combiner l'intégration temporelle précoce et tardive permette d'améliorer les résultats.

Le désavantage des techniques d'intégration temporelle précoce présentées est que ces techniques obligent à ensuite prendre une décision sur un horizon pouvant être relativement long. En utilisant la même idée que celle de l'empilement d'attributs, il est aussi possible d'intégrer l'évolution temporelle du signal au niveau des échantillons en utilisant des attributs contextuels [106] : un échantillon est donc représenté par ses propres attributs, mais aussi par ceux des n échantillons précédents et des n suivants.

Si $\mathbf{X_t}$ représente les attributs de l'échantillon s_t, alors la représentation contextuelle de s_t est $\mathbf{X_t}' = \{\mathbf{X_{t-n}}, \ldots, \mathbf{X_t}, \ldots, \mathbf{X_{t+n}}\}$.

1.2.4 Intégration multimodale

Les contenus multimédias sont, en général, constitués de plusieurs canaux d'information, tels que l'image, l'audio ou les sous-titres. L'objectif de l'intégration multimodale est de fusionner l'information disponible provenant des différents canaux d'information, par exemple pour améliorer les performances du système considéré. Considérer les différents canaux d'information disponibles dans un contenu fait surgir différents problèmes, tels que des problèmes de synchronisation des informations entre les différentes modalités. Par exemple, un commentateur de football prononce en général le mot "But !" après que l'évènement soit arrivé. L'utilisation de la l'intégration multimodale pose également des problèmes d'échantillonnage ou encore des problèmes de représentation des informations. C'est un sujet de recherche très actif. Certains articles sont consacrés à une revue de l'état de l'art disponible dans ce domaine, tels que [131, 119, 14]. La revue la plus récente est, à notre connaissance, celle réalisée par Atrey *et al.* [3]. Nous nous appuyons sur cet article pour définir tout d'abord ce qu'est une modalité, puis présenter très brièvement les problèmes liés au niveau de fusion de l'information, à la façon de fusionner les modalités et de les synchroniser ; Enfin, nous citons quelques articles de l'état de l'art ayant trait aux contenus multimédias.

1.2.4.1 Modalité : définition

Lorsqu'on s'intéresse à la multimodalité, il est important de comprendre ce que l'on entend par modalité. Une modalité est un canal d'information spécifique. Dans le cadre des contenus multimédias, cela correspond traditionnellement à l'audio, les images ou les sous-titres par exemple. Cela peut aussi inclure des métadonnées externes, telles que des informations sur le contenu (genre, acteurs, réalisateur, script ou résumé pour un film par exemple). Dans le cadre de contenus vidéo 3D, on peut aussi parler de multimodalité lorsqu'on distingue explicitement les différents flux vidéo provenant des différentes caméras. Dans un système de vidéo surveillance contenant plusieurs caméras, le caractère multimodal du système peut venir de l'exploitation des différents flux visuels pour détecter des évènements. Si l'on dispose d'un système de transcription de la parole, on peut aussi considérer la parole ainsi retranscrite comme une modalité supplémentaire, différente du canal audio.

La réelle difficulté consiste ici à définir et sélectionner ce que l'on souhaite fusionner, car les différentes modalités contiennent en général des informations bien différentes. Atrey *et al.* [3] indiquent que la recherche automatique de l'information pertinente se base essentiellement sur les méthodes de sélection d'attributs, que nous ne décrivons pas ici, ou sur des méthodes de sélection de modalités pertinentes, très peu étudiées dans la littérature. On peut tout de même citer Wu *et al.* [133] qui essayent de trouver des modalités statistiquement indépendantes directement à partir des attributs de façon à maximiser les sources d'information. Ils utilisent ensuite des super-noyaux pour trouver

la combinaison optimale de ces modalités afin de classer des images et ou de reconnaître des concepts vidéos.

1.2.4.2 Fusion de modalités

De même que pour l'intégration temporelle, il existe plusieurs niveaux de fusion de l'information : la fusion précoce et la fusion tardive. Nous considérons dans cette section que les modalités sont synchronisées ou qu'une synchronisation a été effectué en amont.

La fusion précoce consiste principalement à fusionner les vecteurs d'attributs en un seul. Ce type de fusion a deux principaux avantages : il y a un seul modèle à apprendre, et ce modèle peut directement prendre en compte les corrélations entre les modalités au niveau des attributs. En revanche, il peut se poser le problème de la représentation des attributs, qui peut ne pas être la même en fonction des modalités, comme par exemple entre la modalité texte et les modalités audio et vidéo. Il est aussi nécessaire de synchroniser les attributs provenant de différentes modalités, en utilisant par exemple l'intégration temporelle des attributs.

La fusion tardive consiste, quant à elle, à fusionner les décisions prises par des modèles distincts utilisant un sous-ensemble de modalités. Les décisions ont en général la même représentation, ce qui rend plus facile leur fusion que pour la fusion précoce. De plus la fusion tardive offre plus de flexibilité dans le système. En revanche, cette méthode nécessite d'apprendre plusieurs modèles distincts, ce qui peut être long en fonction du modèle et de la quantité de données.

Enfin, Atrey *et al.* [3] mentionnent aussi les techniques hybrides, combinant à la fois fusions tardive et précoce, permettant d'avoir des systèmes modulaires et de profiter des avantages des deux types de fusion. Il est aussi possible d'utiliser les modalités comme des filtres successifs, c'est-à-dire utiliser une modalité pour extraire des segments d'intérêt dans le contenu, puis une autre pour raffiner le résultat.

1.2.4.3 Techniques de fusion tardive

Si la fusion précoce est assez simple à mettre en œuvre dans le cas où les modalités sont synchronisées, la fusion tardive est plus délicate. Atrey *et al.* [3] distinguent trois types de techniques : les méthodes basées règles, les méthodes basées classification et les méthodes basées estimation.

Règles : Ces méthodes utilisent des règles simples, qui peuvent être définies de manière *ad hoc* par exemple. On compte aussi parmi les méthodes basées règles des méthodes telles que la fusion pondérée moyenne ($\hat{D} = \sum_{j=1}^{J} w_j D_j$, où D_j est la décision prise sur la j$^{\text{ème}}$ modalité et w_j le poids qui y est associé) ou encore la fusion par vote majoritaire ($\hat{D}$ est remplacé par la décision la plus présente parmi les décisions prises par les différents classifieurs).

Classification : L'idée est ici d'utiliser les sorties des classifieurs appris sur les différentes modalités comme attributs d'entrée d'un nouveau classifieur, qui va servir pour prendre la décision finale.

Estimation : Ces techniques sont basées sur le suivi temporel des échantillons, au travers de filtres de Kalman ou de filtres à particules par exemple. Ces techniques sont principalement utilisées pour le suivi de trajectoires, de personnes, de locuteur, ...

Il semble que les techniques de fusion tardive les plus utilisées dans le cadre de l'analyse de documents multimédias soient les SVM, les BN et la régression logistique. L'utilisation d'autres techniques de classification, d'estimateurs ou de règles reste minoritaire.

1.2.4.4 Synchronisation de modalités

Dans un document multimédia, il y a deux principaux problèmes liés à la synchronisation des modalités. Tout d'abord, les différentes modalités n'ont pas forcément le même taux d'échantillonnage. Par exemple, si l'on prend le cas d'une vidéo extraite d'un DVD, il y a, en Europe, 25 images et 44 100 échantillons audio par seconde. Il faut donc synchroniser l'extraction des attributs si l'on veut utiliser la fusion précoce, ou synchroniser les décisions si l'on veut utiliser la fusion tardive. Le problème de la synchronisation des modalités est par conséquent intimement lié au problème de l'intégration temporelle. La synchronisation peut par exemple se faire en extrayant les attributs des différentes modalités avec le même taux d'échantillonnage (exemple : fenêtres audio de 40 ms, et attributs vidéo à chaque image), ou en utilisant l'intégration temporelle sur des horizons temporels plus long, comme un plan vidéo. Dans le cas où les taux d'échantillonnage de l'image et de l'audio sont proportionnels, on peut aussi dupliquer les échantillons vidéos ou sous-échantillonner la modalité audio de façon à obtenir le même taux d'échantillonnage pour les deux modalités.

Le deuxième problème, plus délicat à résoudre, vient du fait que la même information n'est pas forcément présente dans les différentes modalités au même moment. Ainsi, un commentateur de sport annoncera une faute après que celle-ci ne se soit produite sur la modalité vidéo. On peut aussi imaginer qu'un évènement généralement multimodal soit présent dans une modalité et pas dans l'autre. Si l'on prend la cas d'un combat dans un film, il est possible d'imaginer une scène où la modalité visuelle montre le combat, sans que celui-ci soit présent dans la modalité audio, ou que la bande sonore soit de la musique. Ce problème est à notre connaissance un problème ouvert de recherche à l'heure actuelle.

1.2.4.5 Multimodalité dans les contenus multimédias

Dans cette section, nous présentons certaines des techniques d'intégration multimodale les plus utilisées dans le cadre des contenus multimédias. Il semble que pour ces derniers, tels que les films, il soit surtout utilisé des méthodes hybrides ou fondées sur la fusion de décisions. Par exemple, Papadopoulos *et al.* [101] étudient les journaux télévisés, en fusionnant les décisions issues d'un ensemble de détecteurs de concepts basés sur des HMM à l'aide d'un BN dont la structure est définie manuellement. On peut aussi citer Giannakopoulos *et al.* [55, 108] qui fusionnent les sorties des détecteurs de

concepts audio et vidéos fondés sur des BN naïfs à l'aide d'un classifieur de type KNN dans le but de détecter la violence dans les films, ou Ionescu *et al.* [61] qui fusionnent les sorties d'un ensemble de détecteurs de concepts utilisant des réseaux de neurones à l'aide d'un autre réseau de neurones, dans le but également de détecter de la violence dans les films. Concernant les méthodes hybrides, on peut citer Chen *et al.* [22] qui ont développé un système audio-visuel utilisant fusion précoce et fusion tardive pour caractériser des scènes de coups de feu. Ils commencent par analyser les sons de coups de feu avec l'audio, les activités humaines dans la scène à l'aide de la vidéo et les émotions en faisant une fusion précoce des attributs audio et vidéo. Les sorties de ces trois modules sont ensuite fusionnées à l'aide d'un classifieur HMM. Citons aussi Jiang *et al.* [66] qui comparent les méthodes de fusion, à différent niveaux, précoce, tardive et hybride, dans le cadre de la détection de violence dans les films. Il semble que la méthode la plus efficace combine fusion précoce et intégration temporelle tardive. Les modalités peuvent aussi servir d'*a priori* l'une pour l'autre, comme dans Yoshitaka *et al.* [136], qui se sont intéressés à la segmentation en scènes d'une vidéo. Ils commencent par détecter les plans vidéos à l'aide de la vidéo, puis ils analysent le son autour des frontières des plans pour détecter les ruptures. Nous classons cette méthode comme hybride, car elle fusionne les décisions du détecteur de plan en s'en servant pour restreindre la fenêtre d'analyse audio aux moments potentiels de changement de scènes. Concernant la fusion précoce, on peut tout de même citer Irie *et al.* [62] qui combinent des dictionnaires de mots audio et de mots visuels pour constituer un dictionnaire audio-visuel pour détecter des émotions dans des films.

1.3 Détection de violence

Nous terminons cet état de l'art par un tour d'horizon de la littérature sur les méthodes de détection de violence, organisé en trois sections. Dans un premier temps, nous présentons les types d'évènements liés à la violence. Nous présentons aussi les contenus les plus cités dans la littérature, puis les modalités les plus utilisées.

1.3.1 Violence dans la littérature

Il semble que la littérature dédiée à la détection de violence ne propose pas de définition générale de la violence. Les auteurs se contentent en général de donner des définitions très précises, applicables à des travaux très spécifiques, comme la violence dans le sport. Aucun de ces travaux ne propose de définition de la violence qui soit plus générale, et qui puisse s'adapter à tous les types de contenus. Par exemple, Chen *et al.* [23] se sont concentrés sur "une série d'actions humaines accompagnées de saignements", Souza *et al.* [40] et Nievas *et al.* [100] sur "des scènes contenant des combats, quels que soient leur contexte et le nombre de personnes impliquées". Quant à Giannakopoulos *et al.* [55], ils se sont intéressés à "des humains qui, intentionnellement, menacent, tentent ou effectivement infligent de la douleur physique à d'autres humains". Pour finir, Gong *et al.* [56] cherchent à détecter "des scènes rapides contenant des explosions, des coups de feu ou des combats entre humains". Ce manque de définition

commune implique directement un manque de jeux de données commun et général, et le faible nombre de jeux de données développés sont dédiés à des évènements violents très spécifiques.

Il faut aussi noter que la notion de violence dans la littérature scientifique semble fortement corrélée avec la notion d'action, ou de combat. Nous présentons donc des articles portant à la fois sur la détection d'actions, de combats et/ou de violence. Par exemple, Chen *et al.* présentent deux systèmes similaires pour détecter les scènes d'action [24], puis pour détecter les scènes violentes [23]. Il semble cependant que la définition de d'action dans [24] et de violence dans [23] soit très similaire. La définition d'action utilisée par Wang *et al.* [130] semble aussi très corrélée à la définition de violence, car elle inclut les scènes de combat et d'explosion, pouvant être liées à la violence.

1.3.2 Contenus ciblés

La détection de violence peut s'appliquer à de nombreux contenus. Par exemple, on peut chercher à détecter les combats dans un système de vidéo surveillance [26], ou dans les vidéos de sport [100], ou encore dans des vidéos synthétiques [37]. On peut aussi imaginer utiliser des vidéos issues de sites de partage de vidéos [40, 139], tel que Dailymotion. Mais le type de contenu le plus visé dans le littérature est sans conteste les films. Par exemple, Vasconcelos *et al.* [127] utilisent 23 bandes annonces de films pour classer les films par genre. Moncrieff *et al.* [96] s'intéressent à la mesure d'affect dans les films en cherchant des motifs particuliers dans le signal audio. Ils utilisent quelques échantillons extraits de deux films. Wang *et al.* [130] utilisent des plans vidéos extraits de cinq films d'action. Giannakopoulos *et al.* [53, 54, 55, 108] utilisent des extraits de films étiquetés violents ou non-violents. Enfin, Lin *et al.* [88, 87] utilisent cinq films divisés en deux parties : une partie test et une partie apprentissage.

On peut globalement noter le manque de base de données commune conséquente et publique dans le cadre des films, ce qui oblige les équipes souhaitant travailler sur le sujet à développer leur propre jeu de données. Cela ne permet évidemment pas de faire des comparaisons objectives entre les différents systèmes proposés.

1.3.3 Modalités utilisées

Il semble que la modalité vidéo soit celle qui rencontre le plus de succès dans la littérature. On peut citer notamment [127, 130, 24] pour les actions, [37, 100, 26] pour les combats ou [127, 40, 23] pour la violence directement. Concernant l'audio, on peut citer Moncrieff *et al.* [96] pour l'affect dans l'audio et Giannakopolous *et al.* pour la détection de violence directement à partir d'attributs [53] ou en passant par un étage intermédiaire de détection de concepts [54].

Il semble en revanche que les techniques récentes aient tendance à utiliser à la fois l'audio et la vidéo. Ainsi, Giannakopoulos *et al.* [55, 108] ont fait évoluer leur système défini dans [54] pour inclure la modalité vidéo. Lin *et al.* [88, 87], Jiang *et al.* [66] et Ionescu *et al.* [61] sont d'autres exemples de techniques multimodales récentes. Citons

tout de même Nam *et al.* [99] comme l'une des premières publications multimodales sur la détection de violence.

Dans l'ensemble, les systèmes de détection de violence suivent le schéma habituel d'indexation de contenu. Les attributs audio et vidéos extraits sont souvent classiques, et les techniques de caractérisation utilisées se limitent généralement aux SVM et aux réseaux bayésiens. Assez peu d'évolutions significatives ont été apportés dans les systèmes par rapport à la littérature sur l'indexation d'évènements. Cela peut être imputé à la nouveauté de l'application, et/ou à la difficulté lié à l'obtention et au partage des données.

Deuxième partie

Détection de concepts audio

Chapitre 2

De la difficulté de la tâche de détection d'évènements dans les films

Ce chapitre a pour but de mettre en avant les problématiques principales liées à l'indexation dans les films. Nous montrons tout d'abord quelles sont les spécificités des films par rapport à d'autres types de contenus tels que les vidéos YouTube. En nous appuyant sur la littérature, nous montrons qu'il y a un *problème de généralisation* lié à la détection de concepts dans les films. Nous constatons ce problème de manière expérimentale, avant de mettre clairement en avant le fait que ce problème peut être lié à une divergence statistique entre les films.

2.1 Présentation de la tâche

Comme indiqué en introduction, nous nous intéressons dans un premier temps à la détection de concepts audio violents, tels que des coups de feu ou des explosions, dans des films.

Ce faisant nous ne nous plaçons pas dans un contexte de classification, mais dans un contexte de détection. Dans le premier cas, il s'agit d'attribuer une étiquette à chaque élément d'une base de documents multimédias, et chaque contenu de la base est considéré dans son ensemble. Dans le second cas, il s'agit de détecter les moments ou les endroits du document qui correspondent à l'évènement ou au concept que l'on souhaite détecter. Cela amène plusieurs difficultés par rapport à la tâche de classification, la principale étant peut être que l'aspect temporel des évènements doit être pris en compte. Là où les systèmes de classification considèrent généralement les documents audio dans leur intégralité, dans un système de détection, il faut commencer par segmenter le flux audio ou vidéo. On peut aussi noter que, dans le cadre des films, les évènements sont souvent très rares, c'est-à-dire peu présents, voire absents de certains documents.

Le son des films fait partie des contenus multimédias les moins étudiés dans la littérature et les raisons en sont nombreuses. Tout d'abord, les films sont un média

payant. Il est donc difficile de mettre en place une base de données, car il faut acheter les films. Il est plus facile de se tourner vers des vidéos libres de droit, que l'on trouve aisément sur internet. Le second point concerne la nature du son extrait des films. Les films étant des œuvres d'art, portant généralement un message subjectif, ils sont sujet à l'interprétation de la personne qui les regardent. La bande son est donc très travaillée, de façon à essayer de produire l'effet escompté sur la personne qui l'écoute. Ainsi, certains sons peuvent être supprimés, amplifiés, ajoutés, modifiés, ... L'édition peut conduire à une grande variabilité de représentations d'un même message dans différents films, et plus grande est la subjectivité du message, plus grande est la variabilité potentielle due à l'interprétation. Le mixage et l'enchaînement des différents évènements audio rendent aussi l'annotation et la tâche particulièrement difficiles. Par exemple, il n'est pas rare dans un film qu'un dialogue soit mélangé avec de la musique dont le volume peut varier au cours du temps, et avec d'autres effets sonores d'ambiance, comme des bruits de voitures.

Le mixage, la variabilité, la rareté des évènements, la difficulté à annoter, tous ces éléments contribuent à la difficulté de la tâche de détection des évènements dans des films.

2.2 Présentation des jeux de données utilisés dans la thèse

Pour illustrer notre propos, nous présentons dans cette section les deux jeux de données ayant servi pour les expériences décrites dans cette partie du mémoire.

2.2.1 Jeu de données préliminaire

Le premier jeu de données utilisé nous a servi à conduire des expériences préliminaires. Comme indiqué dans le tableau 2.1, ce jeu de données comporte 15 films et un extrait d'un documentaire. Trois films ont été utilisés pour les tests, les 12 films et le documentaire restants servant pour l'apprentissage. Les films de ce jeu de données ont été annotés en termes de *cris, coups de feu* et *explosions*.

Cet ensemble comporte une grande quantité de données, plus de 23 heures de films annotées. Il est, de plus, d'une grande diversité. Il contient des films de genres différents, comme *"Saving Private Ryan"* et *"Back To The Future"*, d'époques différentes, comme *"I Am Legend"* et *"Midnight Express"*. On remarque que la durée de cris, coups de feu et explosions dans chaque film est très variable, allant de zéro à plusieurs secondes, le tout totalisant moins de 2% du jeu de données total.

Malgré la grande quantité de données annotées, la personne ayant effectué la tâche d'annotation a utilisé le signal vidéo, annotant les évènements sonores à l'image vidéo près. Devant l'imprécision et les oublis potentiels de cette annotation, il a été décidé de développer un jeu de données supplémentaire que nous décrivons dans la section suivante.

Films	**Durée**			
	Totale	**Coups de feu**	**Explosions**	**Cris**
Données d'apprentissage				
Lascaux	245,64	-	-	-
Back To The Future	6 675,2	46,4	-	12,8
Billy Elliot	6 352,4	-	-	7,52
Eragon	6 424,56	-	35,72	161,76
Harry Potter 5	7 534,92	-	2,12	13,4
I Am Legend	5 777,04	31,72	25,16	239,52
Leon	7 972,8	101,24	24,48	17,28
Midnight Express	6 965,16	-	-	73,52
Pirates 1	8 238,72	23,24	78,36	62,12
Reservoir Dogs	5 712,72	38,04	-	27,24
Saving Private Ryan	9 752,16	363,04	116,32	82,24
The Sixth Sens	6 049,88	2,16	-	39,44
The Wicker Man	5 869,64	-	8,92	26,64
Total	**83 570,84**	**605,84**	**291,08**	**763,48**
Données de test				
Armageddon	8 678,16	27,24	346,12	69,76
The Bourne Identity	6 814,92	23,12	5,04	0,6
The Wizard Of Oz	6 105,8	-	-	4,2
Total	**21 598,88**	**50,36**	**351,16**	**74,56**

TABLEAU 2.1 – Composition du jeu de données préliminaire. Les durées sont données en secondes. Pirates 1 correspond au film *"Pirates of the Caribbean 1 : The Curse of the Black Pearl"*, Harry Potter 5 au film *"Harry Potter and the Order of the Phoenix"* et *"Lascaux"* est un extrait d'un documentaire.

2.2.2 Jeu de données MediaEval Audio (ME-A)

Le deuxième jeu de données a été développé pour tenter de pallier le défaut de validité du premier. Il s'agit de l'ensemble des films de développement de la campagne d'évaluation MediaEval Affect Task 2012 [44], pour lesquels les concepts *coups de feu, explosions et cris* ont été annotés. Nous n'avons pas utilisé l'annotation en cris car cette dernière n'était pas disponible sur l'ensemble des 15 films du jeu de données mais seulement sur neuf d'entre eux. La composition du jeu de données MediaEval Audio (ME-A) est décrite dans le tableau 2.2.

Ce jeu de données contient 15 films que nous avons divisés en deux sous-ensembles : 10 films ont été utilisés pour l'apprentissage et 5 pour les tests. Les films présents dans ce jeu de données sont, à quelques exceptions près, les mêmes que ceux du jeu de données préliminaire. Nous avons annoté les films en utilisant directement la forme

Films	Durée		
	Totale	**Coups de feu**	**Explosions**
Données d'apprentissage			
Billy Elliot	6 349,4	-	-
Eragon	5 985,4	-	25,42
Harry Potter 5	7 953,48	-	139,63
I Am Legend	5 779,88	41,00	27,72
Leon	6 344,52	84,33	13,66
Midnight Express	6 961	15,82	-
Pirates 1	8 239,36	153,41	63,49
Reservoir Dogs	5 712,92	43,34	-
The Sixth Sense	6 178	2,18	-
The Wicker Man	5 870,4	11,84	14,82
Total	**65 374,36**	**351,92**	**284,74**
Données de test			
Armageddon	8 680,12	31,83	496,26
Kill Bill 1	6 370,44	23,36	2,00
Saving Private Ryan	9 750,96	2 501,22	1 229,39
The Bourne Identity	6 816	27,67	5,53
The Wizard of Oz	5 859,2	-	61,95
Total	**37 476,72**	**2 584,08**	**1 795,13**

TABLEAU 2.2 – Composition du jeu de données ME-A en secondes. Pirates 1 correspond au film *"Pirates of the Caribbean 1 : The Curse of the Black Pearl"* et Harry Potter 5 au film *"Harry Potter and the Order of the Phoenix"*.

d'onde du signal audio à l'aide du logiciel WaveSurfer [1]. Ce faisant, il est beaucoup plus facile de repérer précisément le début et la fin des évènements sonores que l'on annote. Comme le montre le tableau 2.2, il y a moins d'oublis, tels que les coups de feu omis dans *"The Wicker Man"* ou *"Midnight Express"* dans le jeu de données préliminaire par exemple. Il est disponible publiquement à l'adresse `https://research.technicolor.com/rennes/vsd/`.

Le tableau 2.2 montre clairement la diversité de présence des évènements dans les films, ainsi que la rareté des évènements considérés.

2.3 Mise en avant de la problématique de généralisation

Dans cette section, nous montrons que la variabilité dans des films entraîne un *problème de généralisation* pour la détection de concepts audio. Cela signifie que si un modèle est appris à l'aide d'un ensemble de films, ce modèle ne sera pas performant

1. Disponible à l'adresse suivante : `http://www.speech.kth.se/wavesurfer/`

sur des nouveaux films. Nous montrons que ce phénomène est dû à la variabilité entre les différents films. Le problème est ignoré dans la littérature, rendant les résultats publiés non représentatifs de la réalité de la tâche. Nous tentons ensuite d'expliquer ce phénomène à l'aide d'une expérience réalisée avec le jeu de données préliminaire.

2.3.1 Dans la littérature

Dans un premier temps, Giannakopoulos *et al.* [54] se sont intéressés à la détection d'évènements audio violents et non-violents dans le but de détecter de la violence dans les films. Ils ont développé un détecteur de concepts pour détecter la présence de musique, de parole, d'autres sons non-violents qui ne sont ni de la parole ni de la musique (classe "autres" dans la suite), de coups de feu, de combats et de cris. En utilisant une combinaison de KNN et de réseaux bayésiens naïfs, ils reportent de bons résultats pour chacune de leurs classes, allant de 63.3% à 85.1% pour le rappel et de 65.6% à 80.3% pour la précision. Cependant, ces bons résultats sont probablement dus au protocole expérimental utilisé. Leur base de données est constituée d'environ 5 000 échantillons audio de 0.5 à 10 secondes, extraits de plus de 30 films, pour un total de 200 minutes environ. Elle est équilibrée en terme de classes, avec approximativement 800 échantillons par classe. Les résultats rapportés sont obtenus à l'aide d'une stratégie de validation *"hold-out"*, séparant les échantillons en différents sous-ensembles de manière aléatoire, sans prendre en compte la provenance de ces derniers. Ainsi, les ensembles d'apprentissage et de test possèdent potentiellement des échantillons provenant des mêmes films. Les résultats rapportés sont donc biaisés par rapport à l'objectif initial de détection d'évènements audio. En effet, l'objectif de cette méthode est de détecter les évènements cibles dans des nouveaux contenus. Si dans l'ensemble de données d'apprentissage et de test, il y a des échantillons dont la provenance est la même, le protocole expérimental n'est plus représentatif des conditions de fonctionnement réel.

Schlüter *et al.* [118, 61] rapportent des résultats en utilisant un protocole expérimental réaliste, confirmant par la même occasion la grande variabilité entre les films. Ils ont participé à la campagne MediaEval Affect Task 2012, pour laquelle ils ont développé un système de détection de violence en utilisant un étage intermédiaire de détection de concepts basé sur des perceptrons multicouches. Ils rapportent les performances de leurs détecteurs de concepts en utilisant une validation croisée de type *"leave-one-movie-out"* sur les 15 films composant le jeu de données ME-A. Un modèle est appris sur 14 films, testé sur le film restant, et ainsi de suite tant que les 15 films n'ont pas servi de test. Cela permet d'obtenir des résultats correspondant à une application réaliste du système. Ils rapportent 14% de rappel et 10% de précision pour les coups de feu, 17% de rappel et 8% de précision pour les explosions [2]. Ces résultats sont en totale contradiction avec les résultats donnés par Giannakopoulos *et al.*, malgré l'utilisation d'une technique état de l'art, et semblent confirmer l'existence d'une grande variabilité entre les différents films.

Les résultats publiés par Trancoso *et al.* [123, 15] tendent aussi vers l'existence de cette variabilité. Leur stratégie est un peu différente des autres : pour pallier la

2. Pour détecter les explosions, les auteurs utilisent à la fois l'audio et la vidéo.

Attributs	Dimension
MFCC	12
Énergie	1
Centroïd fréquentiel	2
Asymétrie spectrale	2
Platitude spectrale	3
TOTAL	**20**

TABLEAU 2.3 – Attributs extraits.

difficulté à obtenir des exemples d'apprentissage, ils ont utilisé une base d'effets sonores spéciaux contenant 47 concepts audio allant du son d'avion au son de dactylographie. Ils ont appris des modèles à base de SVM en utilisant cette base d'effets sonores, qu'ils ont ensuite appliqués sur des films. Les performances qu'ils rapportent sur la base d'effets sonores sont bonnes : en moyenne, la F-mesure atteint 86%, avec un minimum à 45% et un maximum à 99%. En revanche, les performances rapportées sur les films correspondent aux performances obtenues par Schlüter *et al.*.

Ces approches utilisent l'intégration temporelle au travers de l'utilisation de statistiques. Giannakopoulos *et al.* et Trancoso *et al.* utilisent des segments de longueur fixe en agrégeant leurs échantillons sur des segments de 0,5 à 1 seconde. Schlüter *et al.* ont une stratégie légèrement différente : ils augmentent les attributs de leurs échantillons par des statistiques sur 0,5 seconde de contexte autour de l'échantillon.

Ces articles indiquent donc qu'il existe une variabilité entre les différents films, qui altère les performances dans des conditions réelles d'utilisation et provoque un *problème de généralisation*.

2.3.2 De manière empirique

Dans une approche publiée dans la conférence ORASIS [105], nous montrons nous aussi la variabilité entre les différents films. L'expérience mise en place est relativement simple. Nous avons utilisé le jeu de données préliminaire pour détecter cris, coups de feu et explosions dans les films, en utilisant un processus classique.

2.3.2.1 Description du système

Le système que nous avons mis en place se décompose en trois parties : la segmentation du flux audio, l'extraction des attributs et la classification.

Le flux audio est divisé en échantillons de 40 ms, sur lesquels nous extrayons des attributs classiques. Les attributs extraits sont décrits dans le tableau 2.3. Pour chaque échantillon de 40 ms, on totalise ainsi 20 attributs. La durée des évènements qui nous intéresse étant variable, nous choisissons de ne pas intégrer ces échantillons sur le temps à l'aide de statistiques.

Pour la classification, nous utilisons un SVM avec un noyau RBF [3]. Ce choix est mo-

3. Nous avons pour cela utilisé le logiciel LibSVM [20]

Classe	**Nombre d'échantillons**
Cris	6 743
Coups de feu	6 746
Explosions	6 546
Autres	6 747
TOTAL	**26 782**

TABLEAU 2.4 – Nombre d'échantillons par classe dans la base de données d'apprentissage.

tivé par leur grande utilisation dans la littérature, comme indiqué dans la section 1.2.2.3.

Nous avons utilisé le jeu de données préliminaire pour détecter les trois classes cris, coups de feu et explosions. Les échantillons ne correspondant à aucune de ces trois classes ont été réunis dans une classe que nous appelons *autres*.

Nous avons équilibré le nombre d'échantillons de 40 ms de chacune des classes utilisées pour l'apprentissage du modèle. Les échantillons ont été choisis aléatoirement dans chacun des films d'apprentissage (voir tableau 2.1). Pour les trois classes minoritaires coups de feu, cris et explosions, l'extraction des échantillons de la base s'est faite uniquement dans les films contenant ces évènements. Le tableau 2.4 contient le nombre d'échantillons de 40 ms utilisés pour chacune des 4 classes pour apprendre le modèle SVM.

2.3.2.2 Expériences

Pour estimer les performances de notre système, deux protocoles de validation croisée ont été comparés. La validation croisée nous permet aussi de déterminer le jeu de paramètres d'apprentissage permettant d'obtenir le modèle optimal pour la tâche considérée. Notre hypothèse étant que la variabilité inter-films est tellement importante que les résultats en condition réalistes s'en ressentent, nous comparons la validation croisée *aléatoire* ($\mathrm{CV}_{\mathbf{A}}$) et la validation croisée *leave-one-movie-out* ($\mathrm{CV}_{\mathbf{LOMO}}$) :

Validation croisée *aléatoire* : Les sous-ensembles sur lesquels est effectuée la validation croisée sont tirés aléatoirement sur l'ensemble des échantillons composant la base d'apprentissage. Potentiellement, chaque sous-ensemble contient donc des échantillons appartenant à chacun des films. Autrement dit, la provenance des échantillons n'est pas prise en compte. Cette validation croisée est proche du protocole expérimental mis en place par Giannakopoulos *et al.*.

Validation croisée *leave-one-movie-out* : Cette validation croisée prend en compte la provenance des échantillons. Les échantillons de différents films appartiennent à des sous-ensembles différents. Ce protocole expérimental, employé par Schlüter *et al.*, est réaliste dans le sens où, à chaque étape de validation croisée, les modèles sont testés sur des échantillons dont la provenance est différente des données d'apprentissage.

	$\mathbf{CV_A}$		$\mathbf{CV_{LOMO}}$	
	Précision	**Rappel**	**Précision**	**Rappel**
Cris	82,78	82,54	36,56	35,47
Coups de feu	82,12	79,51	35,77	33,84
Explosions	80,46	87,00	21,99	8,91
Autres	83,45	79,75	18,55	73,65

TABLEAU 2.5 – Comparaison des résultats de $\mathrm{CV_A}$ et $\mathrm{CV_{LOMO}}$ en terme de rappel et précision. Les résultats sont donnés en pourcentages.

	Approche $\mathbf{CV_A}$		Approche $\mathbf{CV_{LOMO}}$	
	Précision	**Rappel**	**Précision**	**Rappel**
Cris	0,42	19,30	0,55	23,12
Coups de feu	2,11	24,67	2,44	21,64
Explosions	7,83	11,62	11,68	30,68
Autres	98,73	72,74	99,54	74,99

TABLEAU 2.6 – Résultats sur les films de test du jeu de donnée préliminaire en terme de rappel et précision. Les résultats sont donnés en pourcentages.

Le tableau 2.5 compare les résultats issus des deux types de validation croisée. Ces résultats montrent que l'estimation des résultats est fortement biaisée si l'on ne prend pas en compte la provenance des échantillons. Tant le rappel que la précision des classes cris, coups de feu et explosions passent de $\simeq$80% à $\simeq$30%, voire moins pour les explosions. On vérifie d'ailleurs ces résultats sur les trois films de test du jeu de données préliminaire. Pour chaque type de validation croisée, nous extrayons les paramètres d'apprentissage optimaux que nous utilisons pour apprendre un modèle. Nous appliquons ensuite ce modèle aux échantillons des films de test. Les résultats sont présentés dans le tableau 2.6.

Ces résultats montrent clairement que l'approche prenant en compte la provenance des échantillons, $\mathrm{CV_{LOMO}}$, est la plus réaliste par rapport à la tâche de détection d'évènements dans les films. On remarque que le taux de rappel obtenu sur les films de test est équivalent à celui obtenu avec $\mathrm{CV_{LOMO}}$, quel que soit le type d'apprentissage utilisé. Il y a d'ailleurs peu de différence de résultats sur les films de test entre les deux approches au final, sauf pour les explosions, qui sont beaucoup mieux détectées par le modèle $\mathrm{CV_{LOMO}}$. En revanche, le taux de précision obtenu sur les films de test est bien inférieur à celui obtenu en validation croisée. Cela s'explique assez simplement : d'une part, la validation croisée est appliquée sur une base équilibrée, et d'autre part, sur la base de test, la classe autres est ultra majoritaire. En validation croisée, le faible taux de précision pour les autres indique qu'une partie des échantillons des trois autres classes ont été confondus avec cette classe. Pour ce qui est de la base de test, le rappel est sensiblement le même pour la classe autres, $\simeq$75%, mais les 25% restant, répartis sur les trois autres classes, correspondent à 11-12 fois plus d'échantillons que la quantité totale

d'échantillons dans les trois autres classes. Ainsi, il y a beaucoup plus d'échantillons autres confondus avec les classes cris, coups de feu ou explosions, que d'échantillons dans chacune de ces classes, même si cela correspond à un faible pourcentage de la classe autres. Cela explique donc le faible taux de précision des trois classes de test. De même les échantillons des trois autres classes confondus avec la classe autres n'ont pas d'influence particulière sur le taux de précision de cette classe qui reste très élevé.

La comparaison entre $\mathrm{CV}_{\mathbf{A}}$ et $\mathrm{CV}_{\mathbf{LOMO}}$ met en évidence un problème de généralisation sur des nouvelle données de notre système. Elle met aussi en évidence l'importance d'utiliser un protocole de validation croisée qui soit réaliste et adapté à l'utilisation finale.

2.3.2.3 Diagnostic

Pour tenter d'aller plus loin dans l'explication du problème et identifier des axes de recherche, nous avons utilisé nos modèles sur quatre films dont les échantillons ont servi à la constitution de la base d'apprentissage. Nous souhaitons ainsi vérifier que les modèles obtenus sont bien capables de donner de bons résultats sur des films dont certains échantillons ont servi pendant leur apprentissage. Ensuite, pour tenter d'expliquer les résultats, nous avons calculé la divergence de Jensen-Shannon [89] entre les échantillons de chaque classe et les échantillons de la classe correspondante dans la base d'apprentissage pour chacun des sept films [4]. Pour les films ayant servi à la constitution de la base d'apprentissage, nous pensons que la faible quantité d'échantillons *autres* inclus dans la base d'apprentissage n'aura pas beaucoup d'influence sur la divergence. La divergence de Jensen-Shannon (D_{JS}) est une version bornée et symétrisée de la divergence de Kullback-Leibler (D_{KL}). La divergence de Kullback-Leibler est définie par :

$$D_{KL}\left[P(\mathbf{x})||Q(\mathbf{x})\right] = \sum_n p_n \frac{p_n}{q_n} \tag{2.1}$$

où p_n et q_n sont les probabilités des échantillons de l'ensemble d'apprentissage $P(\mathbf{x})$ et de l'ensemble de test $Q(\mathbf{x})$. En définissant $M(\mathbf{x}) = \alpha P(\mathbf{x}) + \beta Q(\mathbf{x})$, la divergence de Jensen-Shannon est définie par :

$$\begin{aligned} D_{JS}\left[P(\mathbf{x}), Q(\mathbf{x})\right] = \alpha \cdot D_{KL}\left[P(\mathbf{x})||M(\mathbf{x})\right] \\ + \beta \cdot D_{KL}\left[Q(\mathbf{x})||M(\mathbf{x})\right] \end{aligned} \tag{2.2}$$

De plus, si $\alpha = \beta = 0.5$, alors $D_{JS}\left[P(\mathbf{x}), Q(\mathbf{x})\right] \in [0, \ln 2]$. Ainsi, le résultat de la divergence de Jensen-Shannon est facile à interpréter : si $D_{JS} = 0$ alors les deux distributions sont parfaitement identiques, si $D_{JS} = \ln 2$, alors elles sont complètement divergentes.

La figure 2.1 met en relation le rappel obtenu pour chacun des évènements dans chacun des films avec la divergence de Jensen-Shannon correspondante. Comme le montre la courbe de tendance, il y a une corrélation assez nette entre un faible rappel observé sur les films de test et une forte divergence de Jensen-Shannon.

4. Les trois films de la base de test et les quatre films de la base d'apprentissage.

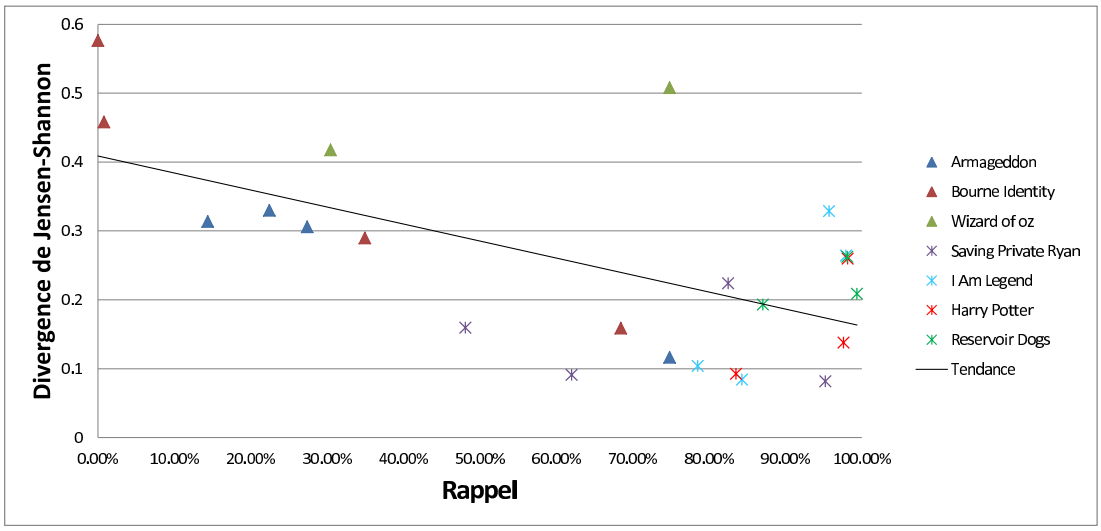


FIGURE 2.1 – Divergence de Jensen-Shannon entre les échantillons des films et ceux de la base d'apprentissage. Les croix correspondent aux films de l'ensemble d'apprentissage, les triangles aux films de l'ensemble de test. Pour chaque film, il y a un point par classe présente.

La figure 2.1 semble donc indiquer que le problème de généralisation est dû à une divergence statistique des attributs entre les différents films. De plus, le point aberrant principal, dont la position est (74,87% ; 0,51), correspond à la classe autres du film de test *"The Wizard of Oz"*. Nous pensons que la classe autres est très polymorphe étant donné qu'elle est composée de tout ce qui n'est pas cris, coups de feu ou explosions, et que, par conséquent, elle occupe tout l'espace statistique entre les classes. La classe autres deviendrait donc un choix par défaut du classifieur, ce qui intensifierait le problème de généralisation.

2.4 Conclusions

Dans ce chapitre, nous expliquons quelles sont les principales difficultés liées au fait de travailler avec des films pour faire de l'indexation. Les films sont un média généralement payant, long et plus complexe que d'autres médias et donc plus fastidieux l'acquisition et l'annotation des données. Sujet à l'interprétation, un même évènement peut être représenté de diverses manières en fonction du réalisateur, ce qui amène une grande variabilité entre les films. De plus, nous travaillons dans un contexte de détection et non pas de classification, nécessitant de détecter des instants temporels de début et de fin des évènements, ce qui rend la tâche plus compliquée.

Nous montrons ensuite que travailler sur des films conduit à un problème de généralisation, et nous montrons par l'expérience que ce problème se manifeste par une divergence statistique des échantillons entre les films. Nous pensons que ce problème peut être dû à deux causes principales : la quantité d'exemples dans la base n'est peut-être

pas suffisante, ou la représentation des évènements est dépendante du contenu lui-même. Nous montrons aussi l'intérêt de prendre en compte la provenance des échantillons dans l'estimation des performances d'un classifieur par validation croisée.

La rareté des échantillons des classes qui nous intéressent est aussi un problème : le déséquilibre que cela entraîne entre les classes affecte grandement les taux de précision des classes minoritaires, et cela même si les taux de rappel peuvent sembler corrects.

Dans notre expérience, nous avons choisi de ne pas prendre en compte l'aspect temporel de la bande sonore des films car nous avons des évènements de tailles variables. Même si les références abordées dans la sous-section 2.3.1 utilisent l'intégration temporelle, nous pensons qu'il est judicieux de bien réfléchir sur ce point, ce qui sera fait dans le chapitre suivant.

Pour conclure, la détection d'évènements audio dans les films se heurte à trois problèmes principaux : *la rareté des évènements*, qui peut affecter la qualité du classifieur, *l'aspect temporel du flux*, qui doit être correctement pris en compte dans le modèle, et enfin, *le problème de généralisation* dû à la divergence statistique entre les films. Dans le chapitre suivant, nous proposons diverses pistes de résolution pour ces différents problèmes.

Chapitre 3

Éléments de résolution des problématiques soulevées

Ce chapitre a pour objectif de présenter les pistes de résolution des différents problèmes que nous avons identifiés dans le chapitre 2, à savoir *la rareté des évènements* et *leur déséquilibre*, *l'aspect temporel du système*, et surtout *le problème de généralisation*. Différentes approches sont présentées dans ce chapitre. Nous présentons tout d'abord une approche ayant pour objectif de prendre en compte la rareté et le déséquilibre des évènements pendant l'apprentissage du modèle de classification. Nos deux autres approches ont pour objectif de réduire la variabilité des attributs en utilisant, dans un premier temps, une approche fondée sur la similarité entre les échantillons, puis, dans un second temps, un approche modélisant la variabilité des attributs.

3.1 SVM à paramètres pondérés

Bien que les expériences décrites dans la section 2.3.2 permettent de montrer le problème de généralisation, elles sont peu réalistes, notamment parce que la base d'apprentissage utilisée pour apprendre les modèles SVM est équilibrée. En effet, le tableau 2.1 montre clairement que, dans les films, la quantité d'échantillons n'est pas équilibrée en terme de classe. Ainsi, en équilibrant la base, le classifieur ne prend pas en compte le déséquilibre existant dans la réalité.

De plus, étant donnée la faible quantité d'échantillons disponibles pour les trois classes qui nous intéressent, l'équilibrage de la base conduit à utiliser un nombre d'échantillons très faible pour la classe autres. Or la classe autres contient tout ce qui n'est pas cris, coups de feu ou explosions, et est donc très polymorphe, potentiellement beaucoup plus que les trois autres classes. Les échantillons utilisés pour modéliser cette classe ne sont donc pas forcément représentatifs de la classe.

Dans la suite de cette section, nous présentons une série d'expériences ayant pour but de répondre au problème de la rareté des évènements et de leur déséquilibre dans la base d'apprentissage.

Classe	**Nombre d'échantillons**
Cris	6 743
Coups de feu	6 746
Explosions	6 546
Autres	110 018
TOTAL	**130 053**

TABLEAU 3.1 – Nombre d'échantillons par classe dans la base de données d'apprentissage.

3.1.1 Présentation de l'expérience

Nous présentons ici les modifications apportées au système présenté en section 2.3.2, modifications introduites dans le but de prendre en compte la rareté des évènements que nous souhaitons détecter.

Tout d'abord, la base d'apprentissage a été modifiée de façon à prendre en compte le déséquilibre entre les différentes classes, d'une part, et la rareté des évènements d'autre part. Le tableau 2.1, page 41, nous montre que la classe autres correspond à environ 98 % de la base de données totale. Pour retranscrire totalement la rareté des cris, des coups de feu et des explosions dans la réalité, c'est-à-dire avoir une base d'apprentissage contenant 98 % de "autres" et 2 % du reste, en gardant le même nombre d'échantillons cris, coups de feu et explosions que dans le tableau 2.4, il faudrait environ 1 million d'échantillons autres, ce qui rendrait l'apprentissage du modèle SVM très complexe (voir section 3.1.3). Un compromis à $\simeq$15 % a donc été réalisé, ce qui correspond à environ 110 000 échantillons autres. Chacune des classes d'intérêt correspond donc maintenant à environ 5 % de la base d'apprentissage. Si cette base ne reflète pas encore totalement la réalité, elle a cependant le mérite d'être utilisable en pratique et d'introduire un déséquilibre entre les classes. Le tableau 3.1 présente la nouvelle répartition des échantillons.

La notion de déséquilibre a également été introduite dans le classifieur à travers l'utilisation d'une pondération dans les SVM. Cette technique permet de biaiser l'hyperparamètre C contrôlant la tolérance aux erreurs du classifieur pour des problèmes de classification sur des données non équilibrées [20]. Il s'agit tout simplement d'utiliser un paramètre C différent en fonction de la classe de l'échantillon considéré. Mathématiquement, la contrainte 1.10 sur les multiplicateurs de Lagrange du problème d'optimisation dual présenté section 1.2.2.3 devient donc :

$$0 \leq \alpha_i \leq C^{\text{classe}[i]} \tag{3.1}$$

où $C^{\text{classe}[i]} = w^{classe[i]} * C$ correspond à la contrainte appliquée à la classe de l'échantillon i. En pratique, cette contrainte est donc indiquée par un biais w pour chaque classe : plus $w^{classe[i]}$ est élevé, plus on donne d'importance à la classe de l'échantillon i. Pour prendre en compte le déséquilibre des classes dans le classifieur, on peut ainsi donner un poids plus important aux classes les moins présentes dans l'ensemble d'apprentissage, et donc rétablir le déséquilibre.

	w = 10		**w = 100**		**w = 1 000**		**w = 10 000**	
	P	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Cris	12,48	30,25	6,77	48,18	6,06	53,66	6,16	52,35
Coups de feu	18,20	32,05	12,80	38,08	10,05	35,59	10,10	35,68
Explosions	7,61	7,64	6,58	8,62	5,83	7,55	5,42	6,75
Autres	92,14	79,78	93,44	45,39	94,09	32,49	93,85	34,77

TABLEAU 3.2 – Résultats obtenus en fonction du poids appliqué sur les classes minoritaires en cross-validation. Les résultats sont donnés en pourcentages.

En résumé, le déséquilibre et la rareté des évènements sont ainsi pris en compte de deux manières : par l'augmentation du nombre d'échantillons autres dans la base d'apprentissage, et par le biais introduit sur l'hyper-paramètre C dans l'apprentissage du modèle SVM.

3.1.2 Expériences

Afin d'évaluer l'influence du paramètre $w^{classe[i]}$, une validation croisée de type $\text{CV}_{\textbf{LOMO}}$ a été appliquée à l'ensemble d'apprentissage du jeu de données préliminaire pour différentes valeurs de $w^{classe[i]}$. Les résultats pour $w^{classe[i]} \in [10, 100, 1\,000, 10\,000]$ pour $i \in [\text{cris, coups de feu, explosions}]$ et $w^{classe[\text{autres}]} = 1$ sont reportés dans le tableau 3.2.

La première chose que l'on remarque est que l'augmentation de $w^{classe[i]}$ a un certain effet sur le taux de rappel des cris et des autres : entre $w^{classe[i]} = 10$ et $w^{classe[i]} = 100$, le rappel des cris passe de $\simeq$30 % à $\simeq$48 %, tandis que celui des autres est quasiment divisé par deux. Ce phénomène s'amplifie avec la valeur de $w^{classe[i]}$. Les taux de rappel des deux autres classes augmentent légèrement. En revanche si les taux de rappel des trois classes qui nous intéressent tendent à augmenter légèrement, l'effondrement du taux de rappel de la classe autres à un effet néfaste sur les taux de précision, et principalement sur celui des cris et coups de feu. Il se trouve en effet que les échantillons autres sont majoritairement détectés en tant que coups de feu, et surtout cris. Il semble donc que trop pondérer les classes minoritaires ait tendance à trop les favoriser, au détriment de la classe autres. Un poids $w^{classe[i]} = 10$ apparaît donc comme le meilleur compromis, car c'est pour cette expérience que l'on obtient les meilleures précisions pour les classes minoritaires, et le plus fort taux de rappel pour la classe autres.

Si les taux de rappel des classes minoritaires sont sensiblement les mêmes que dans l'expérience précédente (voir section 2.3.2), les taux de précision sont plus faibles. La base utilisée pour cette expérience étant déséquilibrée, il serait logique que ces taux de précision soient plus proches de ceux obtenus avec biais sur les films de test que de ceux obtenus sans biais sur les films de test. Pour le vérifier, le meilleur modèle au sens du rappel moyen obtenu par validation croisée sur l'ensemble des quatre classes a été appliqué sur les films de test. Les résultats ont été reportés dans le tableau 3.3.

Si les taux de précision sont équivalents à ceux obtenus précédemment sans biais (voir tableau 2.6, page 46), les taux de rappel sont, eux, nettement inférieurs, semblant

	Films de test	
	P	**R**
Cris	0.61	6.30
Coups de feu	2.05	13.04
Explosions	9.68	9.88
Autres	98.44	84.77

TABLEAU 3.3 – Résultats obtenus sur les films de test en pourcentages ($w^{classe[i]} = 10$).

infirmer les résultats obtenus par validation croisée. Cela peut indiquer soit que la méthode que nous avons employée pour prendre en compte le déséquilibre des classes n'est pas bonne, soit que ce n'est pas le principal problème générateur de mauvais résultats. Ainsi, dans l'hypothèse où la divergence entre les films est le principal problème, il est possible qu'avoir augmenté sensiblement la quantité d'échantillons dans la base d'apprentissage ait amplifié le phénomène, ce qui expliquerait que les résultats avec biais sur la base de test soient moins bons que ceux obtenus sans biais.

3.1.3 De l'utilisation des SVM pour notre problème

Lors de l'expérience présentée dans cette section, nous avons considérablement augmenté la taille de la base de données d'apprentissage, de manière à prendre en compte le déséquilibre des classes en biaisant l'apprentissage. Mais ce faisant, nous avons négligé les contraintes posées par l'utilisation de SVM pour notre problème.

Comme indiqué précédemment, les SVM ont de multiples avantages. Ils sont beaucoup utilisés, et donnent en général de bons résultats. Bien que complexe, le contexte mathématique qui leur est associé est séduisant, assez intuitif et surtout très efficace. La convergence de l'algorithme utilisé par le logiciel LibSVM, qui est celui que nous avons utilisé, est même démontrée sous certaines conditions [25].

Dans la pratique, les SVM ont quelques problèmes limitant l'intérêt de leur utilisation. Tout d'abord, leur temps d'apprentissage est fortement dépendant du nombre d'échantillons utilisés. En effet, l'équation 1.8 et les contraintes 1.9 et 1.10 montrent que chaque échantillon d'apprentissage x_i est associé à un paramètre α_i qui doit être optimisé. Chang et Lin [20] indiquent que, empiriquement, la complexité algorithmique est potentiellement plus que linéaire en fonction du nombre d'échantillons d'apprentissage, ce qui pose des problèmes quand le jeu de données d'apprentissage en contient un grand nombre. Le temps d'apprentissage est aussi dépendant du nombre d'itérations de l'algorithme d'optimisation nécessaire pour converger, et par conséquent dépendant des données d'apprentissage et des contraintes introduites par les hyper-paramètres de l'algorithme. Le temps d'apprentissage est donc difficilement prévisible.

Les hyper-paramètres posent aussi un problème de taille : ils ont besoin d'être réglés de manière fine. S'il existe des méthodes pour régler la valeur du paramètre γ du noyau RBF, telle que l'utilisation de la matrice de Gram du noyau et de la norme de Frobenius [36], les paramètres sont en général réglés à l'aide d'une grille de

recherche associée à une validation croisée. Ceci, ajouté à un temps d'apprentissage en général long, fait que le réglage des paramètres peut prendre beaucoup de temps. A titre indicatif, le réglage des paramètres pour l'expérience présentée dans cette section a pris plus de deux semaines pour les quatre valeurs de $w^{classe[i]}$ testées. Les expériences ont été menées sur une grille de traitement comportant plus de 250 cœurs de processeurs disponibles. Cela correspond, pour six sous-ensembles et une grille de onze valeurs pour C et onze valeurs pour γ, à $6 \times 121 \times 4 = 2\ 904$ apprentissages de modèles pour régler la valeurs des hyper-paramètres.

Étant donnée la faible quantité de données utilisée pour l'apprentissage par rapport aux données disponibles, il est possible qu'utiliser des méthodes de classification moins dépendantes de la quantité de données d'apprentissage, telles que les réseaux bayésiens, nous permettant ainsi d'utiliser la totalité des données disponibles, soit plus efficace.

3.1.4 Conclusions

Dans cette section, nous avons proposé de réduire le déséquilibre des classes et la rareté des évènements en augmentant la quantité de données pour modéliser la classe autres, puis en biaisant les contraintes appliquées aux différentes classes dans l'apprentissage SVM. Nous montrons que les résultats obtenus sont moins bons que ceux obtenus lors l'expérience de la section 2.3.2, ce qui peut être dû à une amplification du phénomène de divergence entre les films créée par l'augmentation du nombre d'échantillons dans la base d'apprentissage. Ceci montre aussi que la divergence semble être le principal problème à résoudre.

Nous expliquons aussi quelles sont les limitations liées à l'utilisation des SVM dans notre système, du fait de la quantité de données disponibles, et du réglage des hyper-paramètres.

Ainsi, la rareté des évènements et le déséquilibre des classes ne semblent pas être les problèmes prépondérants dans notre système. Nous allons donc dans la suite nous atteler à la résolution du problème de généralisation lié à la divergence statistique des attributs entre les films, en nous focalisant sur le traitement des attributs extraits du signal audio.

3.2 Multiples séquences de mots audio

Dans les sections précédentes, nous avons montré que la variabilité inter-films est potentiellement le problème le plus important auquel nous devons faire face. Nous émettons dans cette section l'hypothèse que la variabilité inter-films est portée par les attributs numériques classiques et nous présentons un système simple que nous avons publié à CBMI [107] basé sur l'utilisation du concept de *mots audio*, permettant de grouper les échantillons audio par similarité pour tenter de réduire l'effet de la variabilité entre les films.

Nous tentons aussi de prendre en compte l'aspect temporel du flux audio, à travers l'utilisation de réseaux bayésiens contextuels. Cela nous permet aussi de prendre en

compte le déséquilibre entre les classes de manière intrinsèque, et de bénéficier d'un classifieur dont le temps d'apprentissage est très court.

Nous étudions enfin l'intérêt de construire un modèle basé sur tous les types d'attributs, ou de construire un modèle par type d'attributs et de fusionner leurs sorties par la suite. Cela revient à une fusion tardive des différent types d'information disponible.

Dans cette section, nous commençons donc par présenter le concept des mots audio, ainsi que les travaux qui nous ont inspirés. Nous présentons ensuite le système que nous avons développé et les résultats obtenus pour la détection de coups de feu et d'explosions dans le jeu de données ME-A.

3.2.1 Le concept des mots audio

Le concept des mots audio provient directement de théories sur l'analyse de texte. L'idée de base est de trouver des mots canoniques permettant de réduire la quantité de mots nécessaires pour décrire les documents textuels et ainsi réduire la variabilité qu'il peut y avoir entre eux. Par exemple, le mot *"called"* et le mot *"calling"* peuvent être représentés par la forme canonique *"call"*, ce qui permet de regrouper ensemble les termes similaires. Appliquée à l'image ou au son, l'idée est de grouper par similarité des échantillons de provenance différente. Très utilisé dans la recherche d'image similaire [64] ou dans la représentation de vidéos par sacs de mots [40, 90], l'emploi de ce concept appliqué à l'audio dans un contexte différent de la représentation par sac de mots est plus récente et n'a, à notre connaissance, jamais été appliqué à la détection de concepts audio dans les bandes sons de films.

La principale difficulté liée à la représentation par mots audio consiste à trouver les mots canoniques que l'on souhaite utiliser, c'est-à-dire constituer un dictionnaire des mots canoniques. Pour cela, on effectue en général une quantification de l'espace des attributs. Par exemple, Chin et Burred [17, 30] proposent un système pour découvrir des motifs audio dans un flux audio. Ils extraient pour cela les attributs MFCC, et comparent trois méthodes de construction de dictionnaire : la factorisation en matrices non-négatives, l'analyse en composantes principales et la quantification k-moyennes. Chaque échantillon du flux audio est ainsi représenté par un index dans le dictionnaire, c'est-à-dire que les attributs des échantillons sont remplacés par l'indice du mot le plus proche dans le dictionnaire. Ils utilisent ensuite la séquence de mots ainsi formée pour trouver des motifs dans la séquence. Traditionnellement, chaque échantillon est associé à un mot, mais dans [30], les auteurs proposent d'y associer les K mots les plus proches de l'échantillon, permettant ainsi de préciser le contexte de l'échantillon et de lever les possibles ambigüités liées à la quantification effectuée. En effet, deux échantillons peuvent être représentés par le même mot sans pour autant être équivalents, et considérer le deuxième, voire le troisième mot le plus proche, ajoute de l'information complémentaire. Pour reprendre notre exemple textuel, en modélisant *"calling"* et *"called"* par *"call"*, on perd l'information temporelle portée par le participe passé et le gérondif. Utiliser plusieurs mots du dictionnaire pour les décrire permettrait par exemple de rajouter cette information de manière directe, par exemple *passé* ou *présent*, ce qui réduirait la variabilité entre les différentes formes de passé et de présent. Malgré tout,

les travaux présentés par Chin et Burred [17, 30] restent préliminaires, dans la mesure où l'évaluation de leur système est restreinte à des sons synthétiques courts ou des enregistrements simples.

L'approche présentée par Kumar *et al.* [79] est celle qui ressemble le plus à la nôtre. Dans cet article, les auteurs utilisent les données de TRECVID 2011 Multimedia Event Detection (MED) pour caractériser des contenus vidéos générés par des utilisateurs [1], ainsi que pour détecter des concepts audio dont les annotations sont fournies avec les données. La construction du dictionnaire est réalisée à l'aide d'un processus itératif fondé sur des N-grammes et des HMM, ce qui prend directement en compte l'aspect temporel du flux. Ils construisent ensuite des histogrammes de mots audio sur des segments audio de $\simeq$10 secondes avec 75 % de recouvrement entre les segments successifs, sur lesquels ils appliquent ensuite des classifieurs de type forêts aléatoires [12]. Les résultats obtenus par Kumar *et al.* sur le contenu MED sont bons en termes de taux de rappel, en revanche, nous pensons que le déséquilibre entre les classes explique des taux de précision assez faibles.

Notre système diffère du système de Kumar *et al.* de plusieurs manières. Tout d'abord, le dictionnaire de mots audio est construit de manière différente. Nous prenons en compte l'aspect temporel différemment de Kumar *et al.* : nous utilisons le concept de contextualité présenté section 1.2.3 pour représenter les échantillons, ce qui permet d'obtenir une segmentation fine du flux, c'est-à-dire prendre une décision par segment, et non par groupe de segments. Le type de contenu audio sur lequel nous appliquons notre algorithme fait aussi partie des différences fondamentales. Nous pensons, comme indiqué par les expériences précédentes, que les films génèrent des difficultés bien différentes de celles générées par les contenus MED, notamment de par l'effort d'édition apporté sur la bande son, résultant en une grande divergence des attributs entre les films. Enfin, nous reprenons l'idée développée par Chin et Burred [17, 30] d'utiliser plusieurs mots audio, en l'adaptant à la construction du dictionnaire de manière à préciser l'information apportée aux différents segments.

3.2.2 Description du système

Notre approche se décompose en trois parties, comme illustré sur la figure 3.1. Pour commencer, le signal audio est découpé en segments audio stationnaires, puis des mots audio sont extraits sur ces segments. Enfin, nous utilisons des réseaux bayésiens pour classifier des séquences de mots audio. Nous présentons en détail les trois parties du système dans la suite de cette section.

3.2.2.1 Segmentation du flux audio

Dans un premier temps, le flux audio est segmenté en segments audio stationnaires de longueur variable en utilisant l'algorithme de divergence forward-backward [2]. La longueur des segments ainsi générés varie de $\simeq$10 ms à plusieurs secondes, avec une moyenne oscillant entre 20 et 30 ms.

1. User generated content

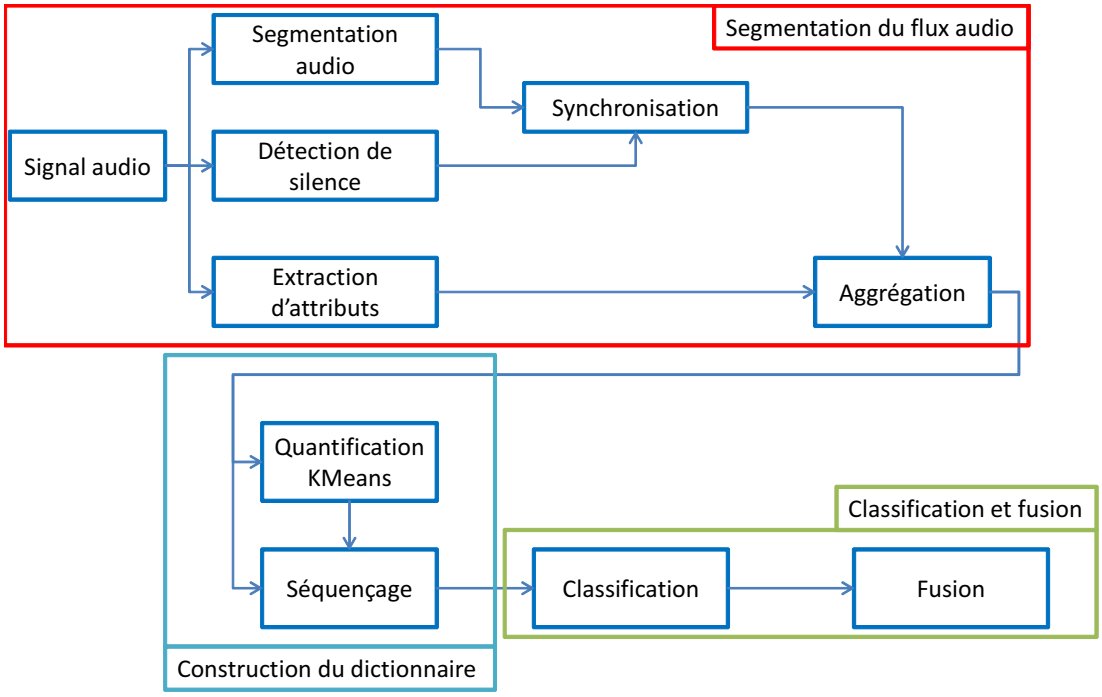


FIGURE 3.1 – Description du système mis en place pour les mots audio.

Parallèlement, les portions silencieuses du signal audio sont détectées en utilisant son profil énergétique. L'histogramme des énergies est approché à l'aide de deux gaussiennes, puis un seuil est obtenu par maximum de vraisemblance pour trouver les segments de basse énergie, correspondant à du silence [2]. La détection des silences est motivée par le fait que les coups de feu et les explosions n'appartiennent pas à cette catégorie. Cette étape constitue donc en quelque sorte un pré-filtrage du signal audio. Les segments audio extraits lors de la première étape sont considérés silencieux si leur intersection avec un segment silencieux est de plus de 50 %. Au final, environ 50 % du nombre total de segments stationnaires est considéré comme silencieux.

Pour finir, des attributs audio sont extraits du signal sur des trames de 20 ms, avec 10 ms de recouvrement. Cela permet de calculer les attributs avec une fréquence de l'ordre de la taille minimale des segments, c'est-à-dire une trame toute les 10 ms. Nous considérons trois types d'attributs classiques :

- **MFCC** : 12 coefficients cepstraux calculés sur l'échelle de Mel.
- **Énergies** : 24 coefficients d'énergie calculés sur un banc de filtres uniforme dans l'échelle de Mel.
- **Platitudes** : 24 coefficients de platitude calculés sur un banc de filtres uniforme dans l'échelle de Mel.

Pour chaque type d'attributs, les dérivées premières et secondes sont calculées, et les attributs sont centrés-réduits film par film. L'agrégation des attributs sur les segments

2. Nous avons utilisé le logiciel AudioSeg, disponible à l'url `gforge.inria.fr/projects/audioseg`

se fait ensuite en faisant la moyenne sur toutes les trames dont le début et/ou la fin se situe à l'intérieur du segment, ou des trames contenant le segment.

Nous obtenons donc en sortie des segments stationnaires de durée variable. A chaque segment est associé un vecteur d'attributs de dimension 36 ou 72 en fonction du type d'attribut utilisé, et une étiquette binaire indiquant s'il est silencieux ou non.

3.2.2.2 Construction du dictionnaire et quantification

L'objectif est maintenant d'associer les attributs audio de chaque segment à un ou plusieurs symboles correspondant aux mots audio. Pour construire le dictionnaire, nous utilisons un algorithme de k-moyennes par quantification par parties [65] [3], qui permet de partitionner l'espace des attributs à l'aide d'un ensemble de vecteurs canoniques appelés moyennes. Le principal inconvénient de l'algorithme des k-moyennes est son coût de calcul sur des ensembles d'échantillons $\mathcal{F} \in \mathbb{R}^D$ de grande dimension D, et de grande cardinalité. Celui-ci peut devenir prohibitif, en particulier si l'on souhaite obtenir un grand nombre de moyennes. La quantification par parties permet de pallier ce problème. Il s'agit d'apprendre des sous-quantifieurs comprenant un nombre plus faible de partitions sur des sous-parties du vecteur d'entrée, puis de recombiner les sorties, de façon à augmenter artificiellement le nombre de partitions. Si l'on considère N sous-quantifieurs avec C moyennes pour chaque quantifieur et si l'on considère une division du vecteur d'attribut d'entrée comme suit :

$$\mathcal{F} := \{\underbrace{X_1, \ldots, X_{\frac{D}{N}}}_{1^{\text{er}}\text{ quantifieur}}, \ldots, \underbrace{X_{D-\frac{D}{N}+1}, \ldots, X_D}_{N^{\text{ième}}\text{ quantifieur}}\} \tag{3.2}$$

alors le nombre de moyennes ainsi réalisées est C^N [65]. Par exemple, apprendre trois sous-quantifieurs avec 128 moyennes permet ainsi d'obtenir 2 097 152 moyennes par combinaison des sorties.

Une conséquence de la quantification par parties est également de permettre la considération des différentes parties du vecteur d'attribut. Par exemple, en suivant l'équation 3.2, si l'on apprend trois sous-quantifieurs sur les MFCC, alors le premier quantifieur correspondra aux coefficients statiques, le second aux premières dérivées, et le troisième aux secondes dérivées. Nous pouvons ainsi considérer un ou plusieurs mots par quantifieur en associant les segments aux différentes moyennes, et obtenir une certaine sémantique ou correspondance sur les différents mots extraits.

Pour résumer, la quantification par parties nous permet d'obtenir N dictionnaires appris sur l'ensemble $\mathcal{F}$ des segments non silencieux pour tous les films d'apprentissage. A chaque segment sont associés K mots pour chaque dictionnaire, choisis comme étant les index des K moyennes les plus proches au sens de la distance euclidienne. Les segments silencieux sont associés K fois à un mot additionnel pour chaque sous-quantifieur. Chaque segment est donc associé au final à KN mots, K pour chacune des N séquences, comme présenté dans la figure 3.2.

3. *Product quantisation* à l'aide des librairies Yael et LibPQ.

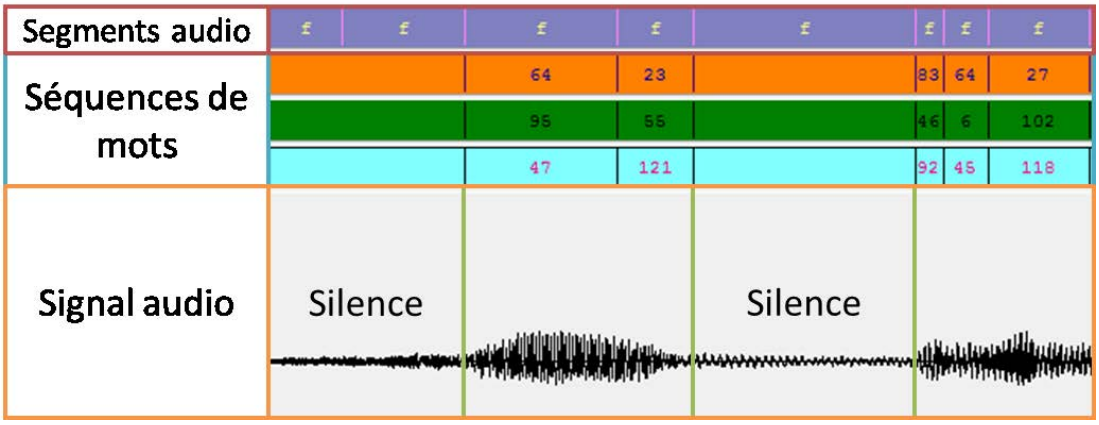


FIGURE 3.2 – Exemple d'une séquence audio après la quantification ($K = 1$ et $N = 3$).

3.2.2.3 Classification et fusion de classifieurs

Un classifieur est ensuite appris sur les mots audio. Nous avons pour cela utilisé les réseaux bayésiens [4]. Ces derniers ont de nombreux avantages par rapport aux SVM : le coût d'apprentissage de leurs paramètres est relativement faible, et il n'y a pas d'hyper-paramètres nécessitant d'être réglés. Au contraire des SVM, le nombre de paramètres des réseaux bayésiens est seulement dépendant de la structure de leur graphe, et en l'absence de variables latentes, ils sont appris par simple comptage des combinaisons dans la base d'apprentissage. En revanche, la complexité de l'inférence augmente rapidement avec le nombre de variables utilisées.

De manière à étudier l'influence de la structure sur le réseau bayésien, nous avons testé deux types de structures de réseaux : la structure naïve et la structure naïve augmentée par une forêt (FAN, plus adaptée que la structure naïve dans un contexte de classification [92].

Nous avons utilisé les échantillons de manière contextuelle de manière à prendre en compte l'aspect temporel du signal, comme suggéré par [69] et [106]. Chaque segment est donc représenté par son contexte sur une fenêtre glissante, c'est-à-dire qu'il est représenté par ses propres mots et les mots des n segments précédents et des n segments suivants.

Nous obtenons donc au final, pour chaque échantillon, la probabilité que ce dernier appartienne à chacune des classes, c'est-à-dire la probabilité qu'il appartienne à la classe coups de feu, qu'il appartienne à la classe explosions et qu'il appartienne à la classe autres. La combinaison des résultats des classes coups de feu et explosions permet, de plus, d'analyser la confusion que le modèle fait entre ces deux classes, c'est-à-dire à quel point le système a tendance à classer les coups de feu comme explosions ou vice-versa.

Enfin, nous étudions la différence entre apprendre un classifieur avec tous les types d'attributs comme entrée (fusion précoce), c'est-à-dire appris sur un vecteur contenant

4. Nous avons utilisé le logiciel Matlab BNT, développé par Kevin Murphy [97], et diponible à l'url `https://code.google.com/p/bnt/`.

les mots extraits des MFCC, des énergies et des platitudes, et la fusion de classifieurs construits sur chaque type d'attributs (fusion tardive). Pour la fusion tardive, nous comparons la fusion par moyenne des sorties des classifieurs (fusion moyenne) et la fusion par somme pondérée des sorties des classifieurs (fusion par poids optimaux), les poids utilisés étant les poids optimaux minimisant l'erreur faite par les classifieurs [80].

3.2.3 Expériences

Comme indiqué en introduction de cette section, nous avons conduit nos expériences à l'aide du jeu de données ME-A, et par conséquent appliqué ce système à la détection coups de feu et d'explosions. Dans un premier temps, nous présentons l'influence des différents paramètres du système par validation croisée sur l'ensemble des 10 films d'apprentissage, puis nous présentons les résultats sur les films de test.

3.2.3.1 Étude sur les différents paramètres du système

Par souci d'obtenir des résultats réalistes, nous étudions les différents paramètres de notre système par validation croisée de type $\text{CV}_{\textbf{LOMO}}$.

Du fait du nombre important de paramètres dans le système ainsi que pour des raisons de lisibilité des résultats, rapporter les résultats quantitatifs détaillés de toutes les expériences permettant d'étudier l'influence des paramètres du système ne serait pas très informatif ni très lisible. Nous nous limitons par conséquent à une discussion basée sur ces résultats. Lors de l'étude de l'influence de chaque paramètre, les autres paramètres restent fixés aux valeurs par défaut suivantes : utilisation des MFCC, structure naïve pour le réseau bayésien, nombre de moyennes $C = 128$, nombre de sous-quantifieurs $N = 3$, nombre de mots extraits par sous-quantifieurs $K = 1$ et profondeur de contexte $n = 5$. Les résultats quantitatifs sont ensuite présentés pour le jeu de paramètres que nous avons jugé comme le plus intéressant.

Structure du graphe : Bien que la structure naïve augmentée de type FAN soit en général plus adaptée que la structure naïve simple dans un contexte de classification, nous observons que l'apprentissage de structure ne semble pas fonctionner dans notre cas. Le réseau FAN ne détecte presque aucun échantillon en tant que coups de feu ou explosions (rappel $< 1\,\%$ pour chacune de ces classes). La structure produite ne semble en revanche pas influencer les taux de précision.

Type d'attribut utilisé : Les attributs que nous avons utilisés sont complémentaires en fonction de la classe. Il semble que les MFCC soient plus efficaces pour les coups de feu (rappel $> 70\,\%$) que pour les explosions (rappel $< 9\,\%$) tandis que les platitudes ou les énergies semblent plus à même de détecter les explosions (rappel $> 50\,\%$) que les coups de feu (rappel $< 20\,\%$). Il semble de plus qu'en utilisant ces trois types d'attributs tous ensemble, c'est-à-dire par fusion précoce, nous obtenions des taux de rappel raisonnables pour ces deux classes ($\simeq$62 % pour les coups de feu et $\simeq$45 % pour

les explosions, voir figure 3.3). Le type d'attributs ne semble pas influencer les taux de précision.

Nombre de mots dans le dictionnaire : Le nombre de mots dans le dictionnaire C est choisi parmi les valeurs $64, 128, 512, 2048$. Sans surprise, il apparaît que plus C est grand, plus le rappel diminue et plus la précision augmente. En effet, comme la quantité de combinaisons possible augmente très vite avec le nombre de mots par sous-quantifieurs, la probabilité de trouver une combinaison inconnue dans l'ensemble de test augmente, ce qui diminue le rappel. Réciproquement, la probabilité que seule une petite quantité de combinaisons ne corresponde qu'à une seule classe augmente, ce qui accroît la précision.

Nombre de sous-quantifieurs et de mots extraits par sous-quantifieurs : Bien que l'augmentation du nombre de sous-quantifieurs ($N \in [1, 3, 9]$) ou l'augmentation du nombre de mots extraits par sous-quantifieurs ($K \in [1, 3]$) accroisse le nombre de combinaisons possibles et devrait par conséquent avoir le même effet que l'augmentation du nombre de mots dans le dictionnaire C, nous observons que le rappel a tendance à augmenter et la précision a tendance à baisser légèrement. Nous pensons que ces paramètres améliorent les capacités de description des mots audio, contrairement à l'augmentation du nombre de mots dans le dictionnaire, et que cela explique le résultat obtenu. En revanche, cela implique l'augmentation du nombre de variables dans le réseau bayésien, ce qui peut rendre la complexité de l'inférence prohibitive si trop de dictionnaires ou trop de mots extraits par dictionnaire sont utilisés.

Profondeur de contexte : Ce que nous appelons contexte se rapporte à la taille de la fenêtre glissante utilisée. Nous avons testé $n \in [2, 5, 10]$, la taille de la fenêtre étant égale à $2n + 1$. Nous avons observé que quand n augmente, le rappel augmente légèrement, mais que la précision baisse, et que la complexité de l'inférence augmente beaucoup.

De manière générale, les taux de précision restent assez faibles (<5-6 %), ce phénomène étant dû au déséquilibre des classes, comme lors des expériences précédentes.

A la lumière de ces expériences, un compromis est réalisé entre rappel et complexité. Les résultats d'un réseau bayésien naïf sur tous les types d'attributs en entrée et avec les paramètres par défaut ($C = 128$, $N = 3$ par type d'attributs, $K = 1$, $n = 5$) sont comparés sur la figure 3.3 aux résultat de Schlüter *et al.* [118, 61], auxquels nous nous référerons dans la suite par "ARF". Les résultats présentés par ARF ont l'avantage d'avoir été produits avec le jeu de données ME-A issu de la campagne MediaEval Affect Task 2012 [44] et donc comparables avec ceux de notre système. Leur système, basé sur l'utilisation de réseaux de neurones de type perceptron multicouches, peut être considéré comme état de l'art. Ils utilisent de plus une validation croisée de type $\mathrm{CV}_{\mathbf{LOMO}}$, mais appliquée à l'ensemble des 15 films (apprentissage + test), tandis que nos résultats de validation croisée sont obtenus sur les 10 films d'apprentissage. Nous obtenons des

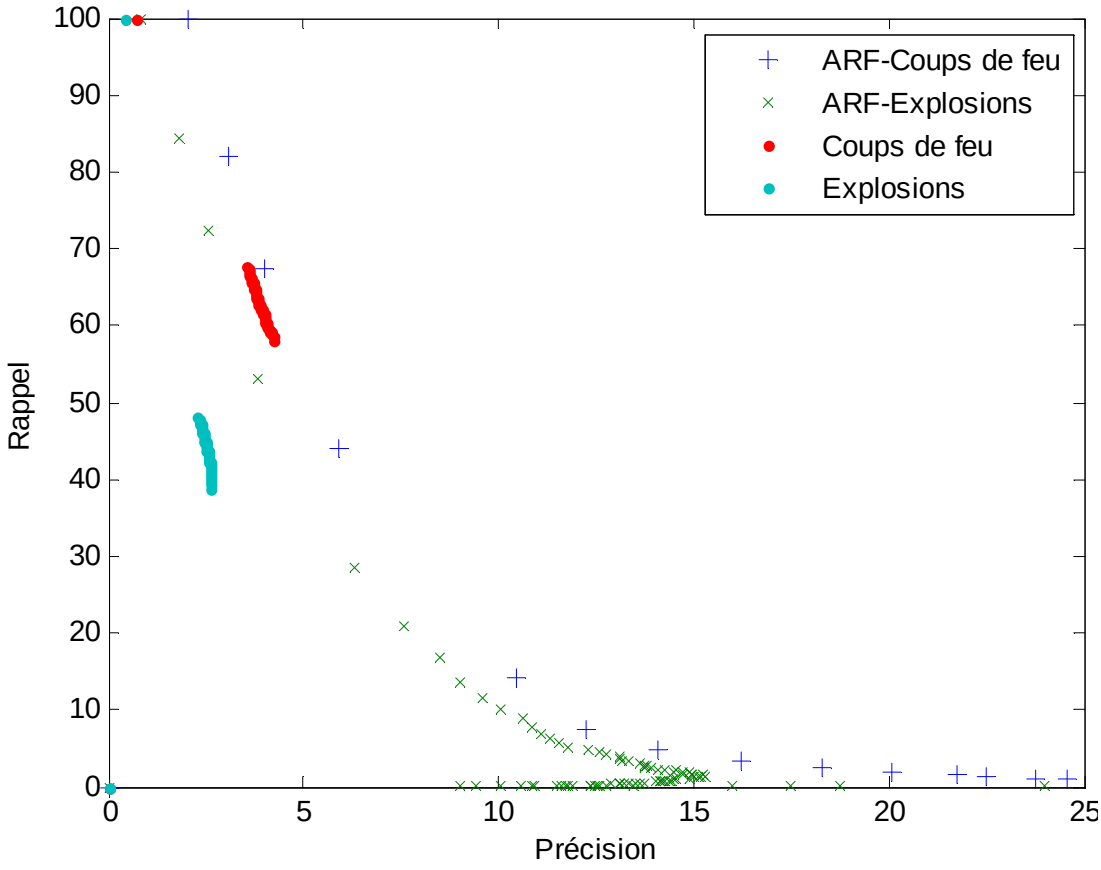
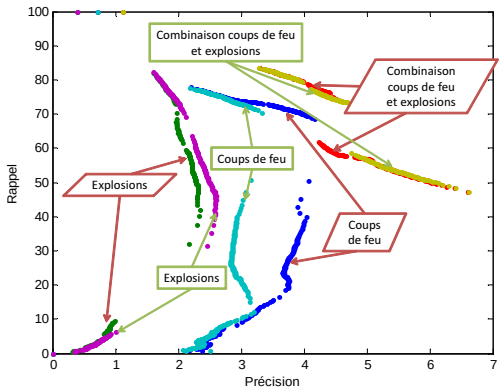


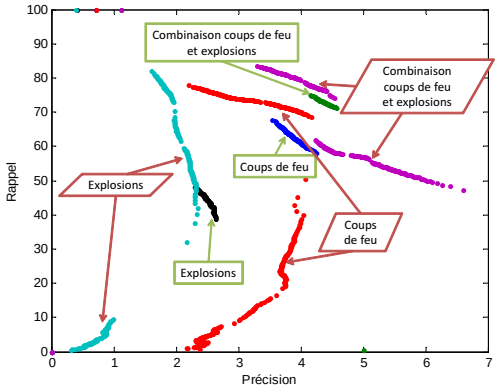
FIGURE 3.3 – Résultats en validation croisée. Comparaison avec les résultats d'ARF.

résultats équivalents aux leurs pour les coups de feu, avec la différence que notre courbe précision/rappel est concentrée dans une zone de rappel élevé. Cette concentration des points indique une décision presque binaire, ce qui démontre la robustesse de notre système. En revanche, bien que concentrée dans une zone de rappel plus élevée, la détection d'explosions est moins bonne dans notre système que pour celui d'ARF. Il faut noter que notre système utilise uniquement l'audio, tandis que celui d'ARF utilise à la fois l'audio et la vidéo pour les explosions, ce qui peut en partie expliquer la différence de résultats.

Nous avons aussi expérimenté la fusion de classifieurs construits à partir de différents types d'attributs. La figure 3.4a compare la fusion moyenne et la fusion par poids optimaux. Les poids optimaux apparaissent comme étant équivalents à la fusion moyenne pour les explosions, et sensiblement meilleurs pour les coups de feu. Malgré cela, la combinaison coups de feu et explosions donne des résultats identiques, ce qui veut dire que la fusion par poids optimaux récupère une partie des coups de feu confondus avec les explosions. De plus, les discontinuités des courbes de la figure 3.4a correspondent principalement à des discontinuités en terme de rappel, et la principale différence entre poids optimaux et moyenne se situe au niveau des seuils de probabilité correspondants.



(a) Fusion moyenne vs. poids optimaux pour la validation croisée. Les rectangles verts correspondent à la fusion moyenne et les parallélogrammes rouges à la fusion par poids optimaux.



(b) Tous les types d'attributs en entrée vs. poids optimaux pour la validation croisée. Les rectangles verts correspondent à tous les types d'attributs en entrée et les parallélogrammes rouges à la fusion par poids optimaux.

FIGURE 3.4 – Résultat des expériences sur la fusion de classifieurs en validation croisée.

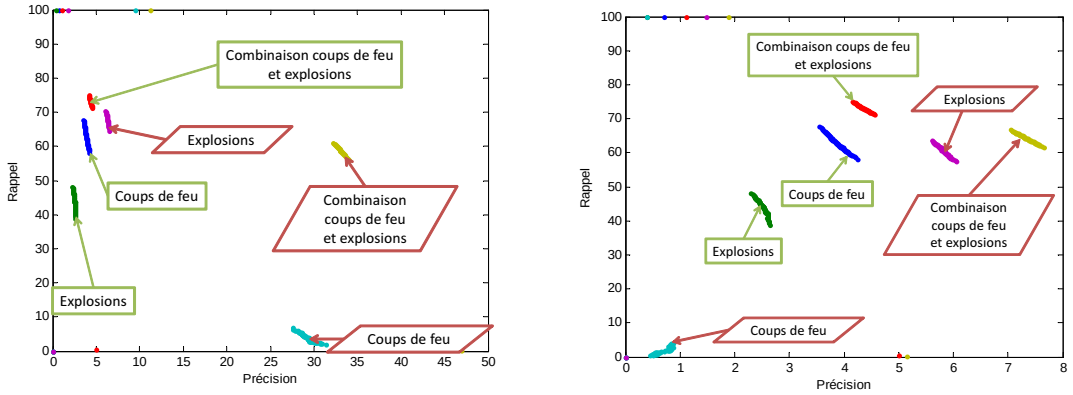
La figure 3.4b compare ensuite les résultats de la fusion par poids optimaux et de la fusion précoce. On remarque que la fusion précoce et la fusion tardive correspondent à des régimes de fonctionnement différents du système : la fusion tardive offre un grand nombre de régimes de fonctionnement possible en fonction du seuil choisi, tandis que la fusion précoce offre une robustesse accrue face au seuil de probabilité en limitant les choix de points de fonctionnement.

Les meilleurs résultats obtenus en combinant les deux classes coups de feu et explosions prouvent qu'il existe une grande confusion entre coups de feu et explosions. Il est intéressant de noter que malgré le déséquilibre important entre les classes, les deux classes qui nous intéressent sont confondues davantage entre elles qu'avec la classe autres.

3.2.3.2 Analyse des résultats expérimentaux

Le modèle issu du compromis réalisé en validation croisée est ensuite appliqué sur les films de test du jeu de données ME-A. Les résultats quantitatifs obtenus pour chaque type d'attributs séparément ne sont pas reportés ici, pour des raisons de lisibilité des résultats, mais restent malgré tout intéressants et contredisent quelque peu les résultats obtenus en validation croisée. Sur les films de test, les MFCC marchent aussi bien que les énergies et les platitudes pour les explosions. Pour les coups de feu, bien que les MFCC rapportent des taux de rappel de l'ordre de 70 % en validation croisée, les taux tombent à presque 0 % sur les films de tests. Nous pensons que ce phénomène est à imputer à la grande variabilité entre les films.

La figure 3.5a compare les résultats de la fusion précoce obtenus par validation



(a) Avec 5 films de test.

(b) Avec 4 films de test (*"Saving Private Ryan"* exclu).

FIGURE 3.5 – Comparaison entre test et validation croisée pour la fusion précoce. Les rectangles verts correspondent à la validation croisée et les parallélogrammes rouges à l'ensemble de test.

croisée et ceux obtenus sur les films de test. La première chose qui apparaît dans ces résultats est le très faible rappel des coups de feu sur les films de test, ce qui est dû à l'utilisation des MFCC. Lorsque l'on s'appuie uniquement sur les deux autres types d'attributs (encore une fois, pour des raisons de lisibilité, les résultats quantitatifs ne sont pas rapportés), le taux rappel des coups de feu obtenu est bien meilleur. En revanche, les résultats rapportés pour la combinaison montre que les coups de feu sont en fait classés en tant qu'explosions, et non en tant que autres. Le deuxième point d'importance est que le rappel des explosions est bien meilleur sur les films de test qu'en validation croisée. Enfin, la dernière chose qu'il convient de mentionner est que les taux de précision obtenus sur les films de tests sont multipliés par 3 pour les explosions et 6 pour les coups de feu comparés aux taux de précision obtenus en validation croisée. Ce résultat est en fait principalement dû à un film dans l'ensemble de test, *"Saving Private Ryan"*, qui contient $\simeq$96 % des coups de feu et $\simeq$70 % des explosions de l'ensemble de test, comme le montre le tableau 2.2. Ainsi, les résultats présentés dans la figure 3.5a sont principalement le fait de ce film, pour lequel les résultats sont particulièrement bons comparés aux autres films. Les résultats ne prenant pas ce film en compte sont présentés dans la figure 3.5b. Les conclusions faites sur les coups de feu et les explosions sont toujours valides, mais les taux de précision obtenus sont bien plus proches de ceux obtenus en validation croisée, atteignant jusqu'à 7-8 %.

Concernant la fusion tardive, nous obtenons les mêmes conclusions que celles obtenues précédemment par validation croisée, à savoir que les poids optimaux semblent légèrement meilleurs que la fusion moyenne.

3.2.4 Conclusions

Toutes les conclusions précédentes indiquent que, bien que les mots audio semblent à même de gérer plus de variabilité que les attributs classiques bas niveau, cette dernière est toujours un problème. Cela se remarque en particulier sur les résultats obtenus sur le film *"Saving Private Ryan"*, ainsi que par la différence entre les résultats des films de test et ceux obtenus par validation croisée. La faible précision obtenue indique qu'il y a toujours une quantité importante de segments audio confondus avec des explosions ou des coups de feu, mais cela correspond à seulement 10-15 % des échantillons autres. Bien que la précision soit toujours faible et que la variabilité soit toujours problématique, les bons taux de rappel sont plutôt encourageants. De plus, nos expériences de validation croisée confirment l'importance de l'utilisation du contexte pour la classification. Nous montrons aussi que l'utilisation de plusieurs séquences de mots audio et/ou de plusieurs mots extraits, i.e., $K, N > 1$, pour décrire les segments améliore les capacités de description des mots audio.

Pour terminer, nous pensons qu'une première étape vers la résolution du problème de généralisation a été franchie à l'aide du concept de mots audio et de l'utilisation de réseaux bayésiens contextuels. Les résultats, obtenus sur le jeu de données ME-A, sont comparables à l'état de l'art établi par ARF sur le même jeu de données, et nous montrons que notre système peut être soit très robuste au seuil, soit proposer plusieurs régimes de fonctionnement. Par la suite, nous introduisons une modélisation de la variabilité de façon à franchir une étape supplémentaire vers la résolution du problème de généralisation.

3.3 Analyse factorielle pour modéliser la variabilité inter-films

Nous nous intéressons dans cette section à la modélisation de la variabilité inter-films à l'origine du problème de généralisation afin, une fois modélisée, de tenter de la compenser dans nos données. Nous utilisons pour cela une technique d'analyse factorielle développée dans le cadre de la reconnaissance de locuteurs permettant de modéliser la variabilité entre les locuteurs et les conditions d'enregistrement des locuteurs directement à partir des données, c'est-à-dire sans connaissance *a priori* sur ces conditions d'enregistrement. Nous avons adapté cette technique à notre problème pour modéliser la variabilité inter-films et nous l'avons implémentée dans notre précédent système.

Dans un premier temps, nous décrivons donc la technique d'analyse factorielle dans son cadre d'origine et présentons l'adaptation du système mise en place. Nous présentons ensuite les résultats obtenus avec cette technique.

3.3.1 Présentation de l'analyse factorielle

3.3.1.1 Contexte

La technique d'analyse factorielle que nous décrivons ici a initialement été développée dans le cadre de la recherche sur la reconnaissance de locuteurs dans des conversations téléphoniques [129]. Cette tâche, au même titre que la nôtre, souffre de problèmes dus à la variabilité entre les locuteurs, mais aussi entre les différentes conditions d'enregistrement de ces locuteurs. L'objectif de la technique consiste donc à modéliser ces différentes variabilités et à s'en servir pour déterminer le locuteur à l'origine d'un nouvel enregistrement. Cette technique a été très utilisée dans le domaine de la reconnaissance de locuteurs, mais aussi pour la reconnaissance de genre vidéo et il en existe de nombreuses variantes [71, 129, 93, 11, 115, 94].

Dans cette section, nous reprenons brièvement les équations décrites par Vogt *et al.* [129], puis nous présentons l'algorithme que nous avons utilisé [93, 11].

3.3.1.2 Principe

Le principe de l'analyse factorielle que nous avons utilisé part de l'hypothèse suivante : à partir d'un modèle GMM représentant le monde (noté UBM [5]), il est possible d'obtenir un modèle adapté à un enregistrement d'un locuteur par modification du super vecteur de moyennes de ce modèle UBM. Cela correspond à une translation du modèle UBM dans un espace vectoriel représentant les locuteurs et dans un espace vectoriel représentant les enregistrements. Dans la suite, par souci de cohérence avec Vogt *et al.*, nous appellerons les enregistrements des sessions.

Le modèle UBM est un modèle GMM à covariance diagonale contenant M gaussiennes de dimension D. Il est appris avec toutes les données utilisables, sans faire de distinctions entre les locuteurs et les sessions. Il s'agit d'un modèle délimitant l'espace des sons disponibles. Ce modèle est défini par son super vecteur de moyennes $\boldsymbol{\mu_{UBM}}$, concaténation des vecteurs moyennes de chacune de ses M gaussiennes, et par $\boldsymbol{\Sigma}$, matrice de taille $MD \times MD$ composée des matrices de covariance $\boldsymbol{\Sigma}_g$ de chacune des gaussiennes sur la diagonale.

En considérant que la variabilité peut venir à la fois des locuteurs et des sessions, le super vecteur de moyennes $\boldsymbol{\mu_{(h,s)}}$ adapté pour la session h du locuteur s peut être trouvé par *maximum a posteriori* (MAP) par :

$$\boldsymbol{\mu}_{(h,s)} = \underbrace{\boldsymbol{\mu_{UBM}}}_{\text{Modèle UBM}} + \underbrace{\overbrace{\mathbf{Dy}(s)}^{\text{Terme adaptation MAP}} + \overbrace{\mathbf{Vx}(s)}^{\text{Terme variabilité locuteur}}}_{\text{Dépendance au locuteur}} + \underbrace{\mathbf{Uz}(h,s)}_{\substack{\text{Dépendance à la session} \\ \text{(et au locuteur)}}}$$

En pratique, le terme de variabilité locuteur n'est pas forcément nécessaire, l'adaptation par MAP $\mathbf{Dy}(s)$ étant souvent suffisante. Seul le terme d'adaptation MAP est

5. De l'anglais *Universal Background Model*.

donc utilisé pour le locuteur :

$$\boldsymbol{\mu}_{(h,s)} = \underbrace{\boldsymbol{\mu}_{UBM}}_{\text{Modèle UBM}} + \underbrace{\overbrace{\mathbf{D}\mathbf{y}(s)}^{\text{Terme adaptation MAP}}}_{\text{Dépendance au locuteur}} + \underbrace{\mathbf{U}\mathbf{z}(h,s)}_{\substack{\text{Dépendance à la session} \\ \text{(et au locuteur)}}} \tag{3.3}$$

Ce modèle adapté à une session d'un locuteur $\boldsymbol{\mu}_{(h,s)}$ est donc constitué de trois composantes : le modèle UBM ; un terme dépendant du locuteur $\mathbf{D}\mathbf{y}(s)$ où $\mathbf{D}$ est la matrice d'adaptation MAP, de taille $MD \times MD$, et $\mathbf{y}(s)$ sont les facteurs du locuteur s ; et un terme dépendant de la session $\mathbf{U}\mathbf{z}(h,s)$ où $\mathbf{U}$ est la matrice de variabilité inter-sessions, de rang faible, de taille $MD \times R_U$, et $\mathbf{z}(h,s)$ sont les facteurs de session, permettant l'adaptation dans le sous-espace vectoriel représenté par $\mathbf{U}$.

Dans la suite, nous présentons brièvement la manière d'estimer les facteurs $\mathbf{z}(h,s)$ et $\mathbf{y}(s)$, ainsi que les matrices $\mathbf{D}$ et $\mathbf{U}$. Nous avons extrait ces équations de Vogt *et al.* [129].

3.3.1.3 Estimation des facteurs locuteurs et sessions

En observant l'équation 3.3, on remarque que les raisonnements menant à l'estimation des facteurs $\mathbf{y}$ et $\mathbf{z}$ sont équivalents. En effet, dans les deux cas, cela revient à estimer un vecteur $\mathbf{q}$ contenant des facteurs permettant de translater un modèle $\boldsymbol{\mu}_{prior}$ dans un espace vectoriel G :

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{prior} + \mathbf{G}\mathbf{q} \tag{3.4}$$

Dans le cas où l'on souhaite estimer $\mathbf{y}$, on fixe $\boldsymbol{\mu}_{prior} = \boldsymbol{\mu}_{UBM}$ et dans le cas où l'on souhaite estime $\mathbf{z}$, on fixe $\boldsymbol{\mu}_{prior} = \boldsymbol{\mu}_{UBM} + \mathbf{D}\mathbf{y}(s)$. Ainsi, $\boldsymbol{\mu}$ est l'adaptation de $\boldsymbol{\mu}_{prior}$ à un jeu d'observations $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_T]$ par translation de $\mathbf{q}$ dans le sous-espace représenté par la matrice $\mathbf{G}$. Pour estimer $\mathbf{q}$, l'estimation par *maximum a posteriori* est utilisée, en fixant la distribution *a priori* de $\mathbf{q}$ comme suivant une loi normale de moyenne nulle et de variance unitaire $\mathcal{N}(\mathbf{q}|\mathbf{0},\mathbf{I})$. Le critère MAP maximise donc :

$$\max_{\mathbf{q}} p(\mathbf{X}|\mathbf{q},\boldsymbol{\mu},\mathbf{G})\mathcal{N}(\mathbf{q}|\mathbf{0},\mathbf{I}) \tag{3.5}$$

Pour résoudre ce problème d'optimisation, un algorithme d'espérance-maximisation (EM) est utilisé. Pour chaque composante g du mélange de gaussiennes, nous définissons l'occupation statistique n_g et la somme pondérée des échantillons $\mathbf{S}_{\mathbf{X},g}$ [6], correspondant respectivement aux statistiques d'ordre zéro et premier de $\mathbf{X}$ pour l'estimation courante de $\boldsymbol{\mu}$:

$$n_g = \sum_{t=1}^{T} P(g|x_t) \tag{3.6}$$

$$\mathbf{S}_{\mathbf{X},g} = \sum_{t=1}^{T} P(g|x_t)x_t \tag{3.7}$$

6. Cela correspond à l'espérance des données $\mathbf{X}$ dans g, multiplié par la cardinalité de $\mathbf{X}$.

Les quantités suivantes sont aussi définies : $\mathbf{S_X}$ est la concaténation $MD \times 1$ des quantités $S_{\mathbf{X},g}$ et $\mathbf{N}$ est une matrice diagonale de taille $MD \times MD$ composée des blocs $\mathbf{N}_g = n_g\mathbf{I_D}$ sur sa diagonale. La statistique de premier ordre est centrée par rapport à $\boldsymbol{\mu_{prior}}$:

$$\mathbf{S_{X|\mu_{prior}}} = \mathbf{S_X} - \mathbf{N}\boldsymbol{\mu_{prior}} \tag{3.8}$$

La solution de l'équation 3.5 est donc donnée par :

$$\begin{aligned} \mathbf{Aq} &= \mathbf{b} && (3.9)\\ \text{avec} & \\ \mathbf{A} &= \mathbf{I} + \mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{NG} && (3.10)\\ \mathbf{b} &= \mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{S_{X|\mu_{prior}}} && (3.11)\\ & && (3.12) \end{aligned}$$

Comme $\mathbf{A}$ est définie positive, on a directement la solution $\mathbf{q} = \mathbf{A}^{-1}\mathbf{b}$. Les preuves de la solution et de la positivité de la matrice $\mathbf{A}$ sont données en annexe A.

Pour le cas particulier des facteurs d'adaptation MAP, $\mathbf{y}(s)$, la matrice D est définie par :

$$\mathbf{I} = \tau\mathbf{D}^T\boldsymbol{\Sigma}^{-1}\mathbf{D} \tag{3.13}$$

où τ est le facteur d'adaptation MAP. On a donc $\mathbf{D} = \frac{\boldsymbol{\Sigma}^{\frac{1}{2}}}{\sqrt{\tau}}$, ce qui nous permet de simplifier $\mathbf{y}(s)$:

$$\mathbf{y}(s) = \frac{\tau}{\tau\mathbf{I} + \mathbf{N}}\mathbf{D}^T\boldsymbol{\Sigma}^{-1}\mathbf{S_{X|\mu_{prior}}} \tag{3.14}$$

Il reste donc à définir la matrice $\mathbf{U}$.

3.3.1.4 Estimation de la matrice de variabilité intersession

Supposons que nous avons S locuteurs, et que $H(s)$ correspond au nombre de sessions disponibles pour le locuteur s. Dans la suite, les paramètres indexés par s, $s \in [1, S]$, seront calculés sur $\mathbf{X}(s) = \{x : x \in s\}$ et ceux indexés par (h, s), $h \in [1, H(s)]$, seront calculés sur $\mathbf{X}(h, s) = \{x : x \in (h, s)\}$.

Alors la matrice $\mathbf{U}$ est définie par :

$$\mathbf{U} = \arg\max_{\mathbf{U}} \prod_{s=1}^{S} p(\lambda_s|\mathbf{X}(1, s), \ldots, \mathbf{X}(H(s), s)) \tag{3.15}$$

où $\lambda_s = \{\mathbf{y}(s), \mathbf{z}(1, s), \ldots, \mathbf{z}(H(s), s)\}$.

En définissant $\mathbf{U}_g$ comme étant la partie de $\mathbf{U}$ calculée à l'aide la $g^{\text{ième}}$ gaussienne, la solution est donnée par :

$$\mathbf{U}_g\mathbf{A}_g = \mathbf{B}_g \tag{3.16}$$

Algorithme 3.1 : Estimation jointe de $\mathbf{U}$, $\mathbf{z}(h,s)$ et $\mathbf{y}(s)$ - Matrouf *et al.* [93]

Initialisation : Pour chaque session h et chaque locuteur s : $\mathbf{y}(s) \leftarrow \mathbf{0}$, $\mathbf{z}(s,h) \leftarrow \mathbf{0}$, $\mathbf{U} \leftarrow$ initialisation aléatoire;
Estimation des statistiques d'ordre zéro et premier : $\mathbf{N}(s)$, $\mathbf{N}(h,s)$, $\mathbf{S}_{\mathbf{X}(s)}$, $\mathbf{S}_{\mathbf{X}(h,s)}$ (équations 3.6 and 3.7);
pour $i = 1$ **à** *itération EM* **faire**
  pour $(h,s) \in (H(s),S)$ **faire**
    Centrer les statistiques de premier ordre : $\mathbf{S}_{\mathbf{X}(s)|\boldsymbol{\mu}_h}$, $\mathbf{S}_{\mathbf{X}(s,h),g|\boldsymbol{\mu}(s)}$ (équation 3.8);
    Estimer $\mathbf{A}(h,s)^{-1}$ et $\mathbf{b}(h,s)$ (équations 3.10 and 3.11);
    Estimer $\mathbf{z}(h,s)$ et $\mathbf{y}(s)$ (équations 3.9 and 3.14);
  fin
  Estimer $\mathbf{U}$ (équations 3.17, 3.18 and 3.16);
fin

où

$$\mathbf{A}_g = \sum_{s=1}^{S} \sum_{h=1}^{H(s)} n_g(h,s) \left(\mathbf{z}(h,s)\mathbf{z}(h,s)^T + \mathbf{A}(h,s)^{-1} \right) \tag{3.17}$$

$$\mathbf{B}_g = \sum_{s=1}^{S} \sum_{h=1}^{H(s)} \mathbf{S}_{\mathbf{X}(s,h),g|\boldsymbol{\mu}(s)} \mathbf{z}(h,s)^T \tag{3.18}$$

La matrice $\mathbf{U}_g$ est donc assez simplement déterminée. La positivité de la matrice $\mathbf{A}_g$ est démontrée en annexe A.

3.3.1.5 Algorithme utilisé pour estimer les facteurs et la matrice

Les paramètres sont estimés conjointement en utilisant l'algorithme 3.1 décrit par Matrouf *et al.* dans [93]. Cet algorithme est semblable à l'algorithme de Gauss-Seidel présenté dans Vogt *et al.* et se base sur l'utilisation d'un algorithme EM. On remarque que les statistiques ne sont calculées qu'une seule fois, au début. Les auteurs de cet algorithme affirment que ce résultat n'a pas d'influence notable sur les résultats, mais que cela accélère la vitesse de l'algorithme.

Il est important de noter que $\mathbf{z}(h,s)$ et $\mathbf{y}(s)$ sont aussi les fruits d'un algorithme itératif, en revanche Vogt *et al.* montrent que si $z(h,s)$ est déterminé en premier, la solution ainsi obtenue après une itération est presque identique à celle obtenue après convergence. Matrouf *et al.* utilisent donc ce résultat et effectuent uniquement une itération pour déterminer $\mathbf{z}(h,s)$ et $\mathbf{y}(s)$.

3.3.1.6 Détermination du locuteur

Une fois les paramètres du modèle calculés, l'objectif est de déterminer quel locuteur s_t est à l'origine du morceau de parole $Y_t = [y_1, \ldots, y_T]$ parmi la collection de locuteurs

$S = [s_1, \ldots, s_S]$ utilisée pour apprendre le modèle d'analyse factorielle. Pour cela la méthode du rapport de la moyenne sur la vraisemblance de chaque échantillon y_i, $1 \leq i \leq T$ (LLR) est utilisée :

$$\begin{aligned} s_t(Y_t) &= \arg\max_{s \in S} LLR(Y_t) \\ &= \arg\max_{s \in S} \frac{LLK(Y_t|\boldsymbol{\mu}_{(h_t,s)})}{LLK(Y_t|\boldsymbol{\mu_{UBM}})} \end{aligned} \tag{3.19}$$

où $LLK(Y_t|\boldsymbol{\mu}_{(h_t,s)})$ correspond à la moyenne des vraisemblances de Y_t par rapport au modèle adapté à la session $h_t = Y_t$ en faisant l'hypothèse que Y_t a été généré par le locuteur s, et $LLK(Y_t|\boldsymbol{\mu_{UBM}})$ correspond à la moyenne des vraisemblances de Y_t par rapport au modèle du monde. Ainsi, Y_t est testé sur chaque locuteur, et celui donnant le ratio maximum est choisi.

Pour éviter de devoir calculer $\mathbf{z}(h_t, s)$ pour chaque locuteur, on fait l'hypothèse que la variabilité due à la session est indépendante du locuteur. Le facteur de variabilité intersession $\mathbf{z}(h_t)$ est donc calculé une fois à partir de $\boldsymbol{\mu_{prior}} = \boldsymbol{\mu_{UBM}}$, et l'effet de la session est directement retiré des échantillons [125] :

$$\hat{Y}_t : \{\hat{y}_i = y_i - \sum_{g=1}^{M} P(x_t|g) \cdot \mathbf{U}_g \mathbf{z}_g(h_t)\} \tag{3.20}$$

Le LLR est ainsi calculé en utilisant $\hat{Y}_t$ à la place de Y_t, et le ratio est fait entre le modèle du locuteur $\boldsymbol{\mu}_s = \boldsymbol{\mu_{UBM}} + Dy(s)$ où $\mathbf{y}(s)$ est donc déterminé à partir de Y_t, et le modèle UBM. Cette méthode est une méthode hybride basée LLR et compensation.

3.3.2 Application à la détection d'évènements dans les films

Il y a plusieurs différences entre notre problème et celui de reconnaissance du locuteur. La première se situe au niveau de la tâche à effectuer : comme son nom l'indique, la reconnaissance de locuteur est une tâche de reconnaissance, c'est-à-dire qu'il faut globalement, dans un fichier son, détecter quel locuteur est à l'origine du morceau de parole étudié. Cela implique que l'on peut considérer le fichier son dans son ensemble sans se préoccuper d'obtenir une segmentation fine de ce dernier, et cela explique que l'on utilise des rapports de moyennes de vraisemblances. Dans notre cas, l'application est moins directe, car nous sommes dans une application de détection des évènements nécessitant une segmentation du flux due au fait que nous ne connaissons pas la durée des évènements que nous cherchons, ni leur emplacement dans le flux. Nous pouvons en revanche supposer que ces évènements sont relativement courts. L'utilisation de rapport de vraisemblance moyenne semble donc difficile dans notre cas.

La deuxième différence principale réside dans les données utilisées. La technique d'analyse factorielle suppose une grande quantité de données dans lesquelles le phénomène de déséquilibre auquel nous faisons face est aussi présent, certains locuteurs parlant moins que d'autres, mais il est sans commune mesure avec le déséquilibre existant dans nos données.

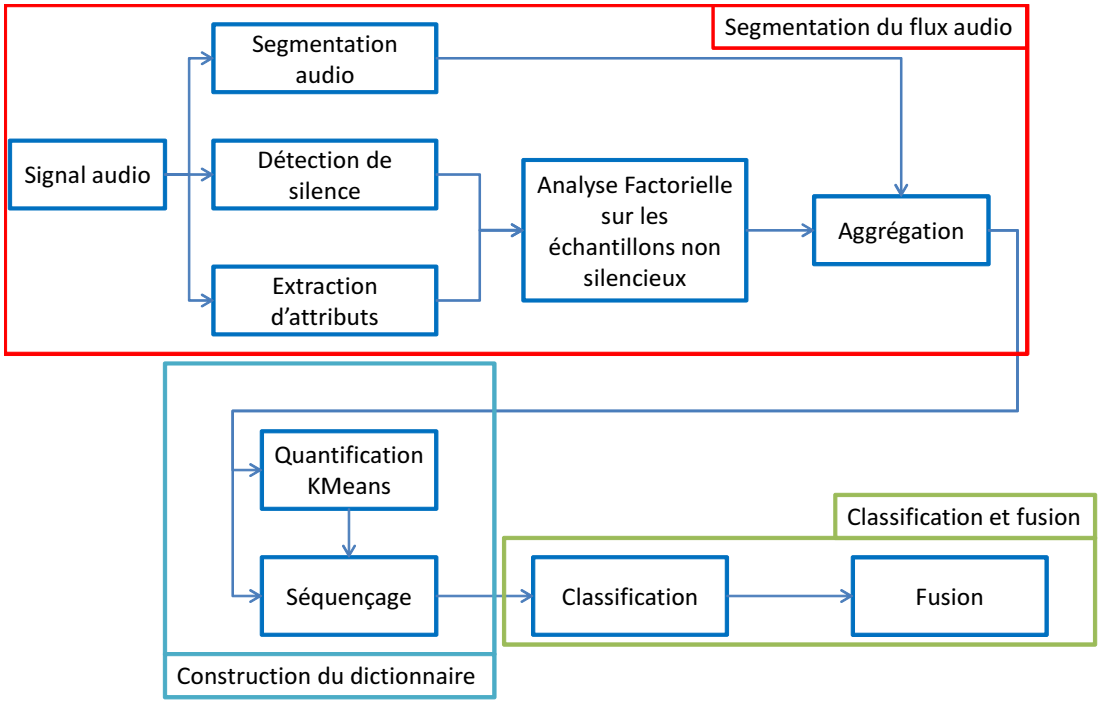


FIGURE 3.6 – Description du système comprenant l'analyse factorielle.

De sorte à ne pas être limité par le problème de la vraisemblance moyenne et par le déséquilibre, nous avons appliqué le problème d'analyse factorielle que nous venons de décrire à la modélisation de la variabilité entre les films. Nous avons pour cela considéré l'analogie suivante avec l'analyse factorielle : l'ensemble des films correspond à un locuteur, et chaque film correspond à une session du locuteur film. Nous avons donc un locuteur et H sessions de ce locuteur, H étant le nombre de films disponibles pour modéliser la variabilité. Une fois cette variabilité modélisée, nous avons utilisé la matrice $\mathbf{U}$ obtenue pour compenser la variabilité dans les attributs extraits des films à l'aide de l'équation 3.20, que nous avons ensuite utilisée dans le système décrit dans la section 3.3.3.

3.3.3 Description du système

Le système que nous avons mis en place est présenté dans la figure 3.6. Il s'agit simplement d'une modification du système proposé à la section 3.2, plus précisément d'une modification de la segmentation du flux audio.

Nous appliquons l'algorithme 3.1 [7] sur les trames de 20 ms non silencieuses des films. Nous nous limitons à ces trames car le reste de l'algorithme se focalise sur ces derniers pour la construction du dictionnaire. Nous pensons de plus que la variabilité entre les

7. Nous avons utilisé les librairies ALIZE et LIA_RAL [11], disponibles gratuitement à l'adresse `http://mistral.univ-avignon.fr/`, pour effectuer nos expériences d'analyse factorielle.

Liste des films ajoutés	
Dead Poets Society	Legally Blond
Fight Club	Misery
Independance Day	Princess Bride
Fantastic Four 1	Psycho
Fargo	Pulp Fiction
Forrest Gump	The Birds
The Pianist	

TABLEAU 3.4 – Liste des films supplémentaires utilisés pour la modélisation de la variabilité.

films est plus présente dans les portions non silencieuses que dans les portions silencieuses. Nous nous limitons aussi à l'utilisation des MFCC car ce sont les attributs pour lesquels cette technique a été développée. Une fois la variabilité calculée, les vecteurs d'attributs sont compensés avec l'équation 3.20, puis agrégés sur les segments audio stationnaires de la même manière que dans le système précédent. Le reste du système est identique.

3.3.4 Expériences

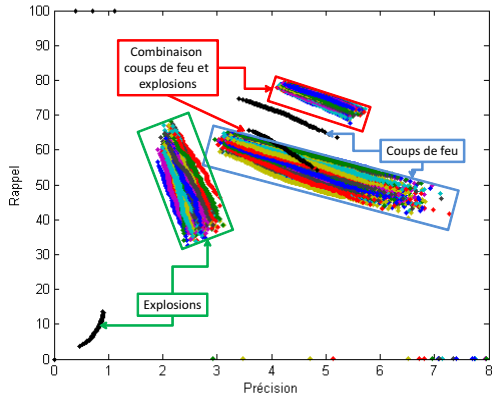
Dans cette section, nous décrivons les expériences que nous avons réalisées et les résultats obtenus grâce à l'introduction de l'analyse factorielle dans notre système.

3.3.4.1 Données utilisées

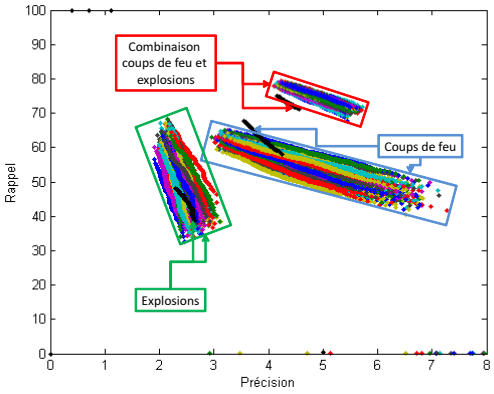
Comme pour les expériences précédentes, nous avons utilisé le jeu de données ME-A décrit dans le tableau 2.2 pour détecter les coups de feu et les explosions. Nous avons en revanche augmenté la taille de ces données pour la modélisation de la variabilité. En effet, les films présents dans le jeu de données ME-A sont les films pour lesquels l'annotation des coups de feu et des explosions est disponible. Comme cette annotation n'est pas nécessaire, il est possible d'ajouter d'autres films dont nous nous servons uniquement pour la modélisation de la variabilité. Nous avons donc rajouté 13 films à la liste des films du jeu de données ME-A, ce qui porte à 28 le nombre de films utilisés, i.e., $H = 28$. Le tableau 3.4 contient la liste des 13 films supplémentaires utilisés pour la modélisation de la matrice $\mathbf{U}$. En revanche, seuls les 15 films du jeu de données ME-A sont compensés, car l'annotation coups de feu et explosions est disponible uniquement pour ces 15 films.

3.3.4.2 Résultats obtenus

Nous proposons tout d'abord d'étudier les différents paramètres du système dérivant de l'utilisation de l'analyse factorielle dans notre système par validation croisée de type $\text{CV}_{\textbf{LOMO}}$.



(a) Comparaison avec les résultats obtenus avec les attributs MFCC.



(b) Comparaison avec les résultats obtenus par fusion précoce.

FIGURE 3.7 – Résultats obtenus pour l'analyse factorielle et comparaison avec les résultats obtenus sans. Les résultats obtenus sans analyse factorielle sont représentés en noir dans les deux figures.

Les paramètres liés à l'analyse factorielle sont le nombre de gaussiennes du modèle UBM M, le rang de la matrice de variabilité R_U et le facteur d'adaptation τ que nous fixons à 16, valeur classique de la littérature. Les valeurs classiques de la littérature pour les deux autres paramètres sont $R_U \simeq 50$ et $M \simeq 1\ 024$, nous avons donc exploré les valeurs suivantes : $R_U \in [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$ et $M \in [8, 16, 32, 64, 128, 256, 512, 1\ 024, 2\ 048, 4\ 096]$.

La première chose que nous avons remarquée est que plus le nombre de gaussiennes dans le modèle UBM augmente, plus il semble que la probabilité que la matrice $\mathbf{A}(h, s)$ soit mal conditionnée après un certain nombre d'itérations de l'algorithme 3.1 augmente, ce qui rend impossible son inversion et donc l'estimation de $\mathbf{U}$. Il semble de plus que les rangs R_U élevés soient les plus touchés. En effet, à partir de $M = 256$, les expériences ont souvent échoué pour cause de mauvais conditionnement de la matrice $\mathbf{A}(h, s)$. Ce résultat entre en contradiction avec la théorie, qui nous indique que la matrice $\mathbf{A}(h, s)$ est inversible. Nous pensons que ceci est dû à une erreur d'approximation numérique mais nous n'avons pas pu le vérifier dans l'algorithme étant donnée la taille des données. Ce problème de mauvais conditionnement des matrices survient soit au niveau de l'apprentissage de la matrice de variabilité, soit au niveau de la compensation des attributs. Sur un total de 100 expériences, 32 expériences ne se sont pas terminées.

La figure 3.7a présente les résultats obtenus en validation croisée sur les expériences qui ont abouties, et les compare aux résultats obtenus précédemment en validation croisée avec les MFCC. Les résultats issus de l'analyse factorielle ont été encadrés par des rectangles de couleur différente en fonction de la classe concernée. Chaque courbe correspond à un jeu de paramètres différents. Nous observons que l'analyse factorielle apporte une nette amélioration aux résultats, principalement pour les explosions. Le rappel obtenu pour ces dernières passe d'un niveau $\simeq$8 % à un niveau compris entre 30

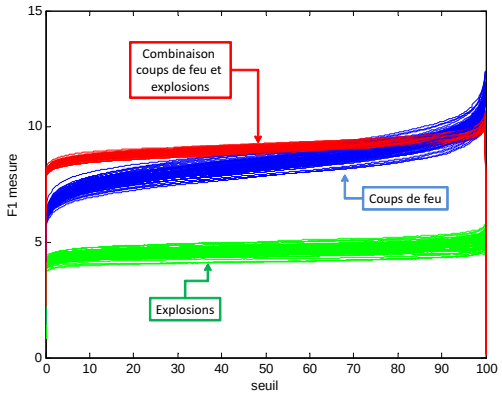
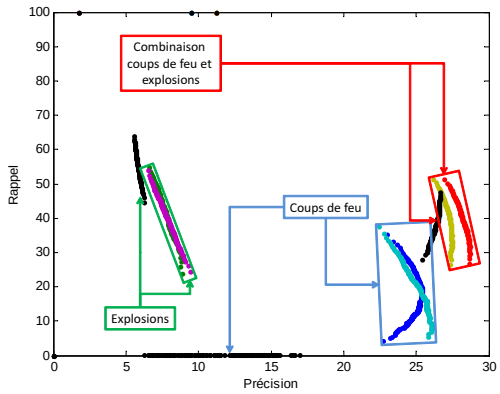


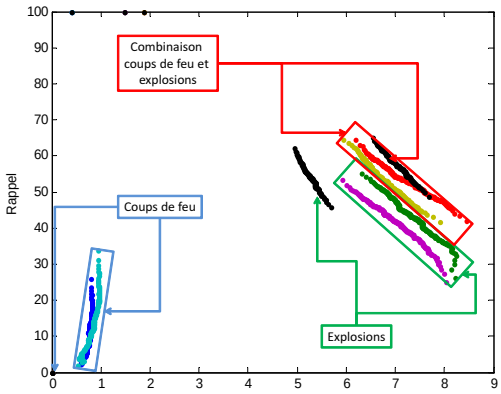
FIGURE 3.8 – F1-mesure obtenue pour les expériences d'analyse factorielle en validation croisée.

et 70 %, et leur précision passe d'un niveau <1 % à un niveau compris entre 2 et 3 %. Avec analyse factorielle, on observe que les résultats obtenus sur les coups de feu sont légèrement dégradés en terme de rappel par rapport aux résultats obtenus sans analyse factorielle. En revanche, en terme de précision, les courbes couvrent une gamme de valeurs plus importante. Enfin, dans le système précédent sans analyse factorielle, on observait que la combinaison des coups de feu et des explosions dégradait légèrement les résultats, ce qui indiquait que pour les MFCC, le système confondait des échantillons explosion avec les échantillons autres. On observe cependant que l'analyse factorielle corrige ce défaut et que la combinaison des coups de feu et des explosions montre cette fois-ci que les explosions et les coups de feu sont plus confondus entre eux qu'avec la classe autres. La figure 3.7b compare ces résultats avec ceux obtenus par fusion précoce en validation croisée sans analyse factorielle. On observe que les résultats obtenus sont équivalents à la fusion précoce, voire légèrement supérieurs pour la combinaison coups de feu et explosions. Cela montre que l'analyse factorielle a un effet bénéfique tellement important qu'on est capable d'obtenir des résultats équivalents à la fusion précoce utilisée dans le système précédent en utilisant moins d'information. Cela montre aussi que l'analyse factorielle a bien comme effet de réduire la variabilité dans les attributs.

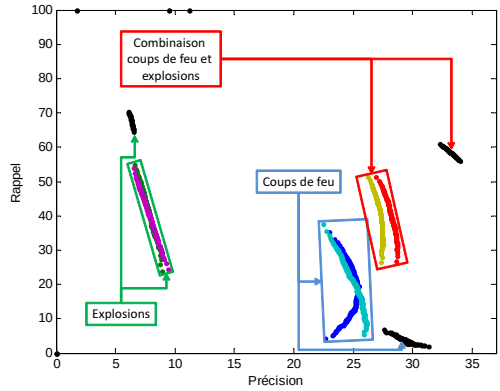
Il semble en revanche que les différents paramètres de l'analyse factorielle n'ont pas d'effet important sur les résultats. Les courbes restent toutes similaires. Malgré une investigation plus poussée, nous n'avons pas réussi à déterminer l'influence des paramètres sur les résultats. Il semble que cette influence soit quelque peu aléatoire. Le choix d'un jeu de paramètres particulier à appliquer sur les films ne semble donc pas avoir de réelle importance. Nous avons tout de même utilisé la mesure F_1, présentée sur la figure 3.8, pour choisir deux jeux de paramètres à tester. Le premier jeu de paramètres correspond aux paramètres donnant la valeur maximale pour la mesure F_1 pour la combinaison des coups de feu et des explosions. Cependant, en observant la figure 3.8, on remarque que le maximum de mesure F_1 correspond à un seuil presque égal à 100 %. Nous avons donc aussi extrait le jeu de paramètres donnant la valeur



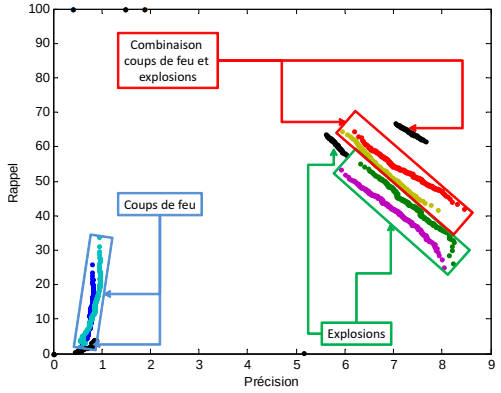
(a) Comparaison avec les résultats obtenus avec les attributs MFCC sur tous les films de test.



(b) Comparaison avec les résultats obtenus avec les attributs MFCC sans *"Saving Private Ryan"*.



(c) Comparaison avec les résultats obtenus par fusion précoce sur tous les films de test.



(d) Comparaison avec les résultats obtenus par fusion précoce sans *"Saving Private Ryan"*.

FIGURE 3.9 – Résultats obtenus pour l'analyse factorielle sur les films de test. Comparaison des résultats obtenus avec analyse factorielle (en couleur dans des rectangles) et sans analyse factorielle (en noir uniquement).

maximale de mesure F_1 pour la combinaison des coups de feu et des explosions à un seuil de 50 %. Les deux jeux de paramètres sont donc : $\theta_1 = \{M = 128, R_U = 10\}$ pour la valeur maximale de mesure F_1, et $\theta_2 = \{M = 1\,024, R_U = 20\}$ pour la valeur maximale de mesure F_1 pour un seuil de 50 %.

Dans les figures 3.9a, 3.9b, 3.9c et 3.9d, nous présentons les résultats obtenus sur les films de test avec les jeux de paramètres θ_1 et θ_2, et nous comparons à chaque fois avec les résultats obtenus sur les MFCC sans analyse factorielle (figures 3.9a et 3.9b) ainsi qu'avec les résultats obtenus par fusion précoce sans analyse factorielle (figures 3.9c et 3.9d). On remarque que les paramètres de l'analyse factorielle n'ont pas un grand effet sur les résultats, ce qui confirme les résultats de la validation croisée. On observait pour les coups de feu sans analyse factorielle que le rappel tombait presque à zero. Il

semble que l'analyse factorielle corrige quelque peu ce défaut : le taux de rappel des coups de feu atteint désormais 40 %. La précision est aussi améliorée, car celle-ci atteint désormais 25 % environ. Concernant les explosions, on remarque que le taux de rappel est légèrement inférieur à celui obtenu sans analyse factorielle, en revanche les taux de précision sont meilleurs, atteignant jusqu'à 10 %. On remarque aussi que l'amélioration du taux de rappel des coups de feu ne conduit pas à une amélioration du taux de rappel de la combinaison mais à une amélioration légère du taux de précision, ce qui signifie que l'analyse factorielle semble améliorer la discrimination entre la classe autres et les classes coups de feu et explosions. Nous faisons les mêmes observations en présentant les résultats sans prendre en compte *"Saving Private Ryan"*, à la différence près que les taux de précision sont beaucoup plus faibles, comme c'est le cas sans analyse factorielle. En revanche, contrairement à la validation croisée, il semble cette fois que les résultats avec analyse factorielle soient moins bons que ceux obtenus par fusion précoce sans analyse factorielle.

3.3.5 Conclusions

Dans cette section, nous présentons une modification du système présenté dans la section 3.2. Dans le but de réduire la divergence entre les films à l'origine du problème de généralisation, nous proposons une adaptation de l'analyse factorielle développée dans le cadre de la reconnaissance du locuteur, où la variabilité dans les conditions d'enregistrements est modélisée par *maximum a posteriori*. Nous utilisons cette technique pour modéliser la variabilité entre les films en utilisant les échantillons non silencieux, puis pour compenser les attributs extraits des films en supprimant la variabilité. Nous avons ensuite utilisé ces attributs dans le système présenté section 3.2.

L'étude des paramètres de l'analyse factorielle par validation croisée montre que ces derniers semblent ne pas avoir d'influence prévisible sur les résultats. Les résultats montrent que quels que soient le jeu de paramètres et la classe considérée, les courbes rappel/précision sont concentrées dans les mêmes zones. Nous montrons que malgré cela, les résultats obtenus améliorent sensiblement les résultats obtenus sur les MFCC sans analyse factorielle, et que les résultats obtenus sont équivalents à ceux obtenus par fusion précoce sans analyse factorielle.

Deux jeux de paramètres choisis à l'aide la mesure F_1 ont été appliqués sur les films de test, et les résultats obtenus montrent qu'il y a bien une amélioration, particulièrement pour les coups de feu, mais que les résultats obtenus par fusion précoce sans analyse factorielle restent meilleurs.

On observe, au travers des différentes expériences menées, que l'adaptation de l'analyse factorielle à notre problème améliore les capacités de généralisation de notre système, même si les résultats ne sont pas parfaits. Il semble, de plus, que les paramètres de l'analyse factorielle n'aient pas une grande influence sur la qualité des résultats, ce qui permet de choisir un couple de paramètres permettant une estimation rapide de la variabilité. Nous en concluons donc que ce type de modélisation de la variabilité est une piste intéressante pour des travaux futurs.

Le chapitre suivant récapitule les contributions que nous avons apportées tout au

long de ce chapitre et présente des perspectives de recherche à nos travaux.

Chapitre 4

Conclusions et perspectives

Nous avons présenté dans cette partie nos travaux ayant trait à la détection d'évènements sonores dans la bande son de films. Ce chapitre sert de conclusion à ces travaux : nous y rappelons nos principales contributions, puis nous présentons des perspectives pour des travaux futurs.

4.1 Contributions

Cette partie du mémoire a été l'occasion de présenter nos expériences sur la détection d'évènements sonores dans les films, en nous concentrant sur la détection de cris, de coups de feu et d'explosions dans un premier temps, puis sur la détection de coups de feu et d'explosions uniquement dans un second temps.

Notre première contribution a été de produire un jeu de données portant sur 15 films Hollywoodiens annotés en termes de coups de feu et en explosions, dont la liste est donnée dans le tableau 2.2. Ce jeu de données, développé dans le cadre de la campagne d'évaluation MediaEval 2012 Affect Task a été rendu public [1].

Nous montrons dans le chapitre 2 que le fait de travailler sur des films amène plusieurs problèmes. Tout d'abord, de par la durée des films, la quantité de données à traiter est énorme et les évènements qui nous intéressent correspondent à une quantité marginale par rapport à la quantité totale de données à traiter. Il faut donc segmenter le flux de manière suffisamment fine pour pouvoir les détecter précisément, ce qui peut poser des problèmes concernant la qualité et l'interprétation des résultats. Cela pose aussi des problèmes de temps d'apprentissage des modèles car cela augmente le nombre d'échantillons disponibles. Ensuite, nous montrons qu'il existe une variabilité telle entre les films que cela conduit à un problème de généralisation du modèle. Nous montrons que ce phénomène de généralisation est présent dans la littérature mais qu'il est ignoré, rendant ainsi les résultats publiés non représentatifs de la tâche. Nous le montrons aussi expérimentalement, à l'aide d'un système simple fondé sur des SVM appris avec une base d'échantillons équilibrée, et en comparant deux types de validation croisée, l'une prenant en compte la provenance des échantillons, l'autre constituant des

1. `https://research.technicolor.com/rennes/vsd/`.

sous-ensembles de manière aléatoire en ignorant la provenance des échantillons. La deuxième méthode donne de très bons résultats, mais très éloignés de ceux obtenus sur de nouvelles données, tandis que la première donne d'emblée des résultats sensiblement équivalents à ceux obtenus de nouveaux les films de test. Nous montrons, toujours expérimentalement, que ce phénomène est certainement dû à une importante divergence statistique entre les attributs des différents films.

Nous explorons ensuite plusieurs pistes pour résoudre ce problème de généralisation. Nous faisons tout d'abord l'hypothèse que ces mauvais résultats sont possiblement dus au déséquilibre entre les évènements. Nous appliquons une pondération par classe sur l'hyper-paramètre C des SVM de sorte à prendre en compte ce déséquilibre pendant l'apprentissage du modèle tout en augmentant la quantité d'échantillons *autres* dans la base d'échantillons d'apprentissage des SVM, et nous montrons que les résultats de généralisation obtenus sont moins bons que ceux des SVM appris sur une base équilibrée. Nous pensons que l'augmentation de la quantité d'échantillons *autres* a amplifié le problème de généralisation.

Dans le but ensuite de s'attaquer à ce problème de généralisation, nous proposons un système fondé sur l'utilisation du concept de mots audio, permettant ainsi de grouper des échantillons de provenances différentes par similarité. Dans le cadre de la détection de coups de feu et d'explosions, nous proposons pour ce faire d'utiliser la quantification par partie, permettant d'extraire plusieurs séquences de mots tout en gardant une sémantique sur les différents mots. Nous proposons de plus une segmentation fine du flux audio combinée à l'utilisation de réseaux bayésiens contextuels, ayant pour avantages d'intégrer l'aspect temporel du signal, et de permettre la prise en compte de la totalité des données pendant l'apprentissage. Les résultats montrent une nette amélioration, notamment en terme de rappel, mais aussi de précision. Les différents types d'attributs utilisés semblent complémentaires, c'est-à-dire donner des performances différentes selon les classes, et nous montrons que la fusion précoce permet d'obtenir un système robuste au seuil de probabilité ayant des performances correctes pour les deux classes coups de feu et explosions, tandis que la fusion tardive par moyenne ou poids optimaux permet d'obtenir un système proposant une grande quantité de points de fonctionnement. Une comparaison avec l'état de l'art montre que les résultats obtenus sont équivalents. Cependant, les expériences effectuées indiquent que la variabilité entre les films n'est pas totalement compensée.

Nous nous sommes intéressés ensuite à la modélisation de cette variabilité, à l'aide d'une technique d'analyse factorielle développée dans le cadre de la reconnaissance de locuteur. Nous proposons alors d'adapter cette technique à la modélisation de la variabilité inter-films, et nous modifions du système précédent de façon à compenser la variabilité ainsi mesurée directement dans les attributs. Appliquée sur les attributs MFCC, cette technique apporte une nette amélioration par rapport au système précédent même si les résultats ne sont pas encore parfaits.

D'une manière générale, nous proposons dans cette partie des briques fondées sur l'utilisation de mots audio pour la résolution du problème de généralisation que nous mettons en lumière. Nous améliorons nettement les résultats par rapport à un système simple, et nous promouvons l'utilisation de réseaux bayésiens contextuels naïfs pour des

problèmes de classification. Nous pensons donc qu'une première étape a été franchie dans la résolution du problème de généralisation, mais que la route à parcourir est toujours longue.

4.2 Perspectives

Les expériences réalisées mettent en lumière diverses pistes techniques pour résoudre le problème de généralisation à l'aide du concept de mots audio. Nous présentons dans cette section celles que nous jugeons intéressantes poste par poste.

Évènements audio : Le jeu de données ME-A a été annoté uniquement en termes de coups de feu et d'explosions, tandis que le jeu de données sur lequel nous avons effectué nos expériences préliminaires intégrait aussi le concept de cris. Il serait intéressant de le rajouter au jeu de données ME-A et d'observer les résultats. D'une manière générale, il serait intéressant de rajouter tout évènement audio auquel on pourrait penser dans le système pour étudier la capacité du système à passer à l'échelle en termes d'évènements.

Attributs audio : Les attributs audio que nous avons utilisés, à savoir les MFCC, les énergies et les platitudes, sont des attributs très classiques, mais ce ne sont pas les seuls. Il est donc possible que d'autres attributs soient plus efficaces ou plus adaptés pour notre problème. Des attributs *ad hoc* peuvent aussi être imaginés, mais cela limiterait l'intérêt du système pour de nouveaux évènements. Enfin, les techniques de projections discriminantes d'attributs, telles que l'analyse discriminante linéaire (LDA), peuvent être envisagées.

Segmentation du signal : Les attributs audio extraits des bandes son des films sont agrégés sur les segments audio par simple moyennage sur la durée des segments. D'autres types de statistiques peuvent être envisagés tels que la variance. On pourrait aussi envisager d'extraire des attributs audio spécifiques sur la durée des segments. La segmentation que nous avons utilisée produit par ailleurs des segments audio de taille variable. Prendre en compte la durée de ces segments dans le système en tant qu'attribut audio pourrait être intéressant. Enfin, un groupement des segments en unités sonores cohérentes à l'aide d'une segmentation hiérarchique permettrait d'obtenir un *a priori* supplémentaire dans le système, pour améliorer la quantification ou la détection.

Création du dictionnaire : Nous trouvons l'idée de quantification temporelle utilisée par Kumar *et al.* [79] séduisante, et nous pensons que cette idée pourrait être appliquée à l'algorithme de quantification par parties, ce qui permettrait de quantifier temporellement les segments.

Classification : La perspective qui nous semble la plus intéressante concernant le réseau bayésien est l'apprentissage de structure. Les résultats que nous avons obtenus

dans cette direction méritent d'être approfondis, car ils sont en contradiction avec les résultats obtenus dans la littérature [106, 57].

Analyse factorielle : Les résultats obtenus à l'aide de l'analyse factorielle montrent que ce type de technique est bénéfique pour les résultats de notre système de détection d'évènements audio. L'application de ce modèle à d'autres attributs que les MFCC, pour tenter d'améliorer globalement les résultats sur les trois types d'attributs utilisés dans le système sans analyse factorielle permettrait probablement d'améliorer les résultats. Nous pouvons envisager d'utiliser d'autres modèles d'analyse factorielle tels que le paradigme des I-Vectors [41], ou développer notre propre modèle, en l'intégrant directement aux réseaux bayésiens ou à la quantification vectorielle. Nous pourrions ainsi intégrer la compensation directement dans le modèle. Enfin nous pourrions aussi envisager d'appliquer l'analyse factorielle en ne modélisant pas la variabilité entre les films mais entre les évènements, ou directement entre les mots extraits de la quantification par parties, en l'intégrant dans un processus itératif avec la quantification vectorielle.

4.3 Sur la course aux résultats

Il est important de se rappeler que l'objectif de cette thèse est de travailler sur la détection de violence dans les films à partir de concepts, ce qui fut une des raisons pour lesquelles nous avons travaillé sur la détection de concepts audio dans les films. À partir de là, nous nous posons la question suivante : est-il nécessaire d'avoir un bon détecteur de concepts audio pour obtenir un bon détecteur de concepts sémantiques ?

Derrière cette question se cache la question de la légitimité de la course aux résultats dans le cadre d'un détecteur de concepts. Une réponse judicieuse à cette question a été apportée récemment par l'équipe ARF [118, 61], justement dans le cadre de la détection de violence pour la campagne d'évaluation MediaEval Affect Task 2012, et il nous paraît important de la présenter. Ainsi que nous le montrons dans le chapitre 5, les films utilisés dans le jeu de données ME-A ont été annotés pour une dizaine de concepts audio ou vidéos en tout, pas seulement en termes de coups de feu ou explosions. L'équipe ARF a utilisé un système basé sur des réseaux de neurones pour construire des détecteurs pour chacun de ces concepts. Les résultats obtenus ne dépassent pas 26 % de mesure F_1, avec une moyenne autour de 14.5 % pour l'ensemble des détecteurs. Ils ont toutefois utilisé les sorties de ces détecteurs comme attributs en entrée d'un détecteur de violence basé lui aussi sur des réseaux de neurones. Les résultats qu'ils obtiennent sont malgré tout très bons, puisqu'ils atteignent 49.94 % de mesure F_1, et obtiennent ainsi les meilleurs résultats de la campagne d'évaluation en termes de MAP@100. Ainsi, ils montrent qu'il n'est pas forcément nécessaire d'obtenir de bons détecteurs de concepts pour obtenir de bons résultats de détection de violence.

Dans la partie suivante, nous nous inspirons de ce résultat pour mettre en place nos propres expériences.

Troisième partie

Détection de violence

Chapitre 5

Avant propos : MediaEval

Nous nous intéressons dans cette partie du mémoire à la détection de concepts sémantiques dans les films à partir de concepts objectifs, avec application à la détection de violence dans les films. Les techniques présentées dans les chapitres suivants ont été développées dans le cadre des campagnes d'évaluation MediaEval 2011 et 2012 [1], au sein desquelles nous avons mis en place une tâche d'évaluation (Affect Task) [43] dédiée à la détection de violence dans les films. Une telle campagne d'évaluation n'existait pas avant 2011. De plus, comme nous l'avons indiqué dans le chapitre 1, il n'existait pas à notre connaissance de base de données publique portant sur la violence et s'appliquant sur des films. Dans ce chapitre avant-propos, nous prenons le temps de présenter la tâche que nous avons mise en place en nous concentrant sur sa version 2012, i.e., nous décrivons le problème, les données utilisées et les métriques d'évaluation.

5.1 Description de la tâche

La détection de violence, et la tâche qui y est associée, a plusieurs applications potentielles, mais le cas d'usage qui a motivé nos travaux de recherche est liée au contrôle parental. Imaginons que nous souhaitions regarder un film avec nos enfants (en bas-âge, environ 8 ans), et que nous choisissions de regarder un film tout public. Il est alors possible pour un jeune enfant de regarder un film tel que *"Kick-Ass"*, classé tout public en France par le CNC, alors qu'il s'agit d'un film contenant beaucoup de scènes pouvant choquer certains enfants de cet âge, notamment des scènes de combats et de torture. La notion de violence et la notation des films sont donc très subjectives, et dépendantes de l'interprétation de la personne qui regarde le film. Il pourrait donc être intéressant pour les parents de pouvoir bénéficier d'une pré-visualisation des scènes les plus violentes des films qu'ils souhaitent regarder, ce qui leur permettrait de décider par eux-mêmes s'il acceptent ou non que leurs enfants les regardent.

1. http ://www.multimediaeval.org/

5.1.1 Définition de la violence

La notion de violence étant une notion très subjective, et fortement dépendante du spectateur, il nous faut donc choisir une définition de la violence qui soit claire et qui puisse servir pour annoter des films. L'organisation mondiale de la santé (OMS) définit la violence comme [1] *"l'utilisation intentionnelle de la force physique ou la puissance, réelle ou sous forme de menace, contre soi-même, une autre personne ou un groupe ou une communauté résultant en ou ayant une forte vraisemblance de résulter en douleur, mort, dégâts psychologiques, malformations, ou privations."*[2]. Elle distingue trois types de violence : auto-infligée, inter-personnelle et collective [78]. Chaque catégorie est ensuite divisée en fonction de la nature de la violence, par exemple, physique, sexuelle, psychologique ou privative. Dans le contexte des films et de la télévision, Kriegel [77] définit la violence comme étant *la force déréglée qui porte atteinte à l'intégrité physique ou psychique pour mettre en cause dans un but de domination ou de destruction l'humanité de l'individu.* Ces deux définitions se focalisent sur les actions intentionnelles et n'incluent par conséquent pas les accidents. Nous pensons que ces derniers peuvent être tout aussi choquants car ils peuvent contenir des scènes d'extrême violence. Dans le but d'inclure ce type de scènes et d'être le plus objectif possible, de façon à faciliter l'annotation des films, nous adoptons la définition suivante :

Définition 5.1. Définition de la violence dans les films pour MediaEval Affect Task
Les évènements violents sont ceux montrant de la violence physique ou des accidents résultant en douleurs ou blessures pour un être humain. Ils sont donc limités à des actions violentes et n'incluent pas la violence verbale ou psychologique. Par ailleurs, il fait qu'il y ait à la fois l'action et le résultat.

Bien que nous ayons tenté d'objectiver la définition de la violence, il y a toujours des cas limites pour lesquels la définition reste inadaptée au cas d'utilisation, ou n'est pas suffisamment objective. La définition peut par exemple conduire à rejeter certaines scènes potentiellement choquantes. Par exemple, une scène contenant des cadavres ensanglantés n'est pas considérée violente par rapport à notre définition car on ne voit pas l'action réalisée, mais uniquement le résultat. Au contraire, quelqu'un giflant quelqu'un d'autre ou quelqu'un boitant et montrant des signes de douleurs en marchant sont des évènements considérés comme violents, car l'action est visible et la douleur manifeste. D'autres évènements définis comme "intention de tuer" n'ont pas été annotés, tels qu'une scène où quelqu'un tire sur quelqu'un d'autre sans le toucher mais avec une intention claire de tuer, car ils ne résultent ni en douleurs, ni en blessures pour un être humain puisque l'action a échouée. D'autres exemples de cas limites sont décrits dans [43].

2. Ceci est une traduction de l'anglais de *"The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation."*

5.1.2 Tâches à effectuer

La campagne MediaEval Affect Task propose deux sous-tâches aux participants. La première, obligatoire, est de classifier les *plans vidéos* en violents ou non violents dans les films, en utilisant une segmentation en plans fournie par les organisateurs. La seconde, facultative, est de détecter les segments vidéos, c'est-à-dire donner les instants de début et de fin des segments violents détectés, sans utiliser la segmentation en plans. Pour cela, les participants peuvent utiliser uniquement la vidéo, le son ou les sous-titres des films, les métadonnées externes (c'est-à-dire toutes les autres sources d'information) n'étant pas autorisées. Nous encourageons ainsi le développement de techniques multimodales basées uniquement sur ce qui est disponible dans un DVD.

5.2 Données utilisées

Comme indiqué dans la section 2.2.2, le jeu de données ME-A a été développé dans le cadre de MediaEval. Les films qui le compose constituent en fait l'ensemble des films d'apprentissage du jeu de données MediaEval présenté dans le tableau 5.1. L'ensemble d'apprentissage est très varié en terme de violence, puisqu'il contient des films très violents tels que *"Kill Bill"*, des films de guerre tels que *"Saving Private Ryan"* ou des films peu violents comme *"The Wizard of Oz"*. Aux films du jeu de données ME-A s'ajoutent trois films de test, *"Dead Poet Society"*, *"Fight Club"* et *"Independence Day"*.

Le tableau 5.1 montre la quantité de données violentes dans les films d'apprentissage et de test, à la fois en terme de segments, et de plans vidéo. Les annotations ont été réalisées par trois personnes, pour s'assurer de leur cohérence. On remarque que contrairement aux évènements audio coups de feu et explosions, tous les films contiennent de la violence. La proportion moyenne de plans violents ($\simeq$10 %) est aussi plus importante que la proportion de coups de feu ou d'explosions dans les films. La différence entre la proportion de plans violents et la proportion de violence en durée que l'on peut remarquer dans le tableau 5.1 s'explique, entre autre, par le fait qu'un plan est considéré comme violent dès qu'il intersecte avec un segment violent, peu importe la durée de cette intersection.

En plus des annotations de violence, des concepts audio et vidéos ont été annotés dans les films. Les participants ont tout loisir de s'en servir pour entraîner des détecteurs de concepts. Les différents concepts sont *la présence de feu, de sang, d'armes à feu, d'armes blanches, de combats, de scènes* "gores" *et de poursuites en voiture* pour la vidéo et *la présence de coups de feu, d'explosions et de cris*[3] pour l'audio.

5.3 Métriques d'évaluation

Pour permettre la comparaison des systèmes développés par les participants, nous avons choisi en 2011 d'utiliser un coût pondéré entre la probabilité de fausses alarmes

3. Uniquement annotés sur 9 films de l'ensemble de données.

Films	**Durée (secondes)**	**Nombre de plans**	**Durée violence (%)**	**Plans violents (%)**
Films d'apprentissage				
Armageddon	8 680,16	3 562	14,03	14,6
Billy Elliot	6 349,44	1 236	5,14	4,21
Eragon	5 985,44	1 663	11,02	16,6
Harry Potter 5	7 953,52	1 891	10,46	13,43
I am Legend	5 779,92	1 547	12,75	20,43
Leon	6 344,56	1 547	4,3	7,24
Midnight Express	6 961,04	1 677	7,28	11,15
Pirates Carib, 1	8 239,4	2 534	11,3	12,47
Reservoir Dogs	5 712,96	856	11,55	12,38
Saving Private Ryan	9 751,0	2 494	12,92	18,81
The Sixth Sense	6 178,04	963	1,34	2,80
The Wicker Man	5 870,44	1 638	8,36	6,72
Kill Bill	5 626,6	1 597	17,4	24,8
The Bourne Identity	5 877,6	1 995	7,5	9,3
The Wizard of Oz	5 415,7	908	5,5	5,0
Total	**100 725,82**	**26 108**	**9,62**	**12,89**
Films de test				
Dead Poet Society	7413,24	1583	0,75	2,15
Fight Club	8005,72	2335	7,61	13,28
Independence Day	8834,32	2652	6,4	13,99
Total	**24 253,28**	**6 570**	**5,07**	**10,88**

TABLEAU 5.1 – Composition du jeu de données développé pour MediaEval 2012.

(P_{fa}) et la probabilité de détections manquées (P_{mi}), que nous avons appelé coût MediaEval (MC) :

$$MC = C_{fa} * P_{fa} + C_{mi} * P_{mi} \tag{5.1}$$

Les poids officiels ont été arbitrairement fixés à $C_{mi} = 10$ et $C_{fa} = 1$ pour refléter le cas d'usage à l'origine de nos travaux de recherche sur la détection de violence dans les films : on souhaite ne surtout pas manquer de scènes violentes. En revanche, quelques fausses alarmes sont acceptables. Cette métrique a l'inconvénient que si un système classe tous les plans des films de test comme violents, il obtient $P_{fa} = 1$ et $P_{fa} = 0$, c'est-à-dire $MC = 1$, ce qui est un très bon score.

Pour pallier les inconvénients du coût MediaEval, qui est biaisé en faveur de valeurs de fausses alarmes élevées, nous avons décidé d'utiliser la métrique *Mean Average Precision* à 100 (MAP@100) comme métrique officielle en 2012. Cette métrique, très utilisée dans des contextes de recherche d'information, tels que l'analyse des résultats rendus par un moteur de recherche, correspond tout de même au cas d'utilisation, car on souhaite montrer les scènes les plus violentes. En effet, s'il y a beaucoup de scènes violentes

dans le film, il paraît inopportun de les montrer toutes, une sélection des plus violentes peut suffire. Si le système permet d'ordonner les résultats retournés par rapport à une valeur de confiance quelconque, le MAP@100 nous donne alors une estimation de la qualité des résultats dans les 100 premières scènes retournées par le système. En se limitant aux 100 premiers plans, on limite aussi le risque de fausses alarmes, car on analyse uniquement ces derniers.

Parallèlement à ces métriques d'évaluation officielles, nous avons aussi calculé le rappel, la précision, la mesure F_1, le taux de fausses alarmes et le taux de détections manquées, ainsi que le MAP@20 et le coût MediaEval en faisant varier les poids pour des soucis d'analyse des résultats.

5.4 Conclusions

L'objectif de chapitre était de présenter la campagne d'évaluation MediaEval Affect Task 2011 et 2012, étant à notre connaissance le premier essai d'évaluation des algorithmes de détection de violence. Nous avons mis en place cette campagne, et développé le jeu de données MediaEval, pour proposer aux équipes de recherche travaillant, ou souhaitant travailler, sur ce sujet un cadre commun d'évaluation pouvant servir de base solide à la comparaison de différents systèmes. La présentation de la campagne nous permet aussi de poser un cadre à nos travaux sur la détection de violence dans les films. Les systèmes que nous présentons dans le chapitre suivant utilisent donc les données du jeu de données MediaEval. Nous présentons certains de ces travaux en détails dans les chapitres suivants puis nous les positionnons par rapport à l'état de l'art en présentant les résultats obtenus en 2012 lors de la campagne MediaEval.

Chapitre 6

Expériences sur la détection de violence

L'objectif de ce chapitre est de présenter trois systèmes de détection de violence que nous avons développés. Les deux premiers sont basés uniquement sur le canal audio et utilisent la représentation par mots audio introduite dans la partie précédente, tandis que le troisième utilise à la fois des attributs audio et vidéos bas niveaux et explore l'intégration multimodale dans le cadre des réseaux bayésiens. Le premier système audio et le système multimodal ont tous les deux été évalués lors de la campagne MediaEval 2012. Nous comparons ensuite ces systèmes, et présentons par la même occasion les résultats de la campagne MediaEval 2012.

6.1 Représentation vectorielle

La représentation par mots audio introduite précédemment nous permet de représenter le signal audio par une séquence de mots audio, de la même manière qu'un texte est constitué d'une suite de mots ou qu'une image peut se représenter comme un ensemble de mots visuels. Ainsi, les représentations par sacs de mots développées dans le cadre de la recherche de documents textuels [31] pour représenter les documents textuels sont beaucoup utilisées à la fois pour la vidéo, au travers des sacs de mots visuels (BoV) (cf. [135, 40]), et pour l'audio, au travers des sacs de mots audio (BoA) (cf. [90, 84]). La représentation par sacs de mots est une représentation vectorielle des documents multimédias (image, fichier son, texte).

Dans le cas de la détection de violence pour la campagne MediaEval, nous souhaitons caractériser les plans vidéos, et nous souhaitons pour cela bénéficier des avantages des représentations par sacs de mots et ainsi vérifier leur efficacité pour le problème qui nous intéresse. Nous appliquons cette représentation aux mots audio extraits du signal audio dans la partie précédente. La représentation vectorielle nous permet aussi d'utiliser les mots audio en tant que "concepts non supervisés". Dans cette section, après avoir présenté brièvement le principe de la représentation par sacs de mots, nous présentons le système simple que nous avons mis en place et les résultats obtenus.

6.1.1 Représentation par sacs de mots

La représentation par sacs de mots consiste à représenter des documents textuels par des histogrammes de mots pondérés. Développé dans le cadre du traitement de documents textuels, ce type de représentation permet par exemple de comparer simplement des documents de taille différente, ou de récupérer le(s) sujet(s) d'un document, en extrayant les mots les plus significatifs. Si elle simplifie la représentation d'un document, la relation temporelle qui peut y être associée est en général perdue par l'histogramme.

Il existe plusieurs types de pondération des mots dans les documents. Nous présentons la représentation TF-IDF [70], qui est probablement la plus utilisée et que nous utilisons dans notre système.

Représentation TF-IDF :

La représentation "simple" d'un document consiste à le représenter par un histogramme où chaque dimension i représente la fréquence d'apparition du $i^{\text{ème}}$ mot du dictionnaire dans le document considéré. Ainsi, plus un mot est fréquent, plus il y a de chance qu'il soit représentatif du document. Si l'on se place maintenant dans le cadre de la recherche de documents similaires dans une base de documents, un terme ayant une grande fréquence d'apparition dans beaucoup de documents, tels que les termes *"le-la-les"* dans des documents textuels par exemple, sera en fait assez peu discriminant, ou représentatif de la similarité.

La représentation TF-IDF (de l'anglais *Term Frequency - Inverse Document Frequency*) propose donc de pondérer la fréquence des termes dans les documents par l'inverse de leur fréquence d'apparition dans la base de documents [70]. Cela permet de réduire l'importance de mots fréquents dans de nombreux documents, et par conséquent peu discriminants, et d'augmenter l'importance des mots peu fréquents dans la base, et donc possiblement très discriminants.

En pratique, pour chaque document $d_i \in \mathbf{D}$, $\mathbf{D}$ correspondant à la base de documents de taille M, pour chaque terme $t_j \in d_i$, t_j étant défini dans un dictionnaire V de taille C, on calcule un poids :

$$\begin{aligned} tfidf(t_j, d_i) &= tf(t_j, d_i) * idf(t_j) \\ &= \frac{f(t_j, d_i)}{max\{f(t \in V, d_i)\}} * \log_2 \frac{M}{1 + |\{d \in \mathbf{D} : t_j \in d\}|} \end{aligned} \tag{6.1}$$

où $f(t_j, d_i)$ est la fréquence du terme t_j dans le document d_i, $tf(t_j, d_i)$ est une normalisation de $f(t_j, d_i)$, $idf(t_j)$ correspond à l'inverse de la proportion de documents dans lesquels le terme t_j apparaît et $|\{d \in \mathbf{D} : t_j \in d\}|$ correspond au nombre de documents contenant t_j.

6.1.2 Présentation du système

Le système que nous avons mis en place est présenté sur la figure 6.1. Il se décompose en plusieurs parties : une étape de mise en forme du signal, une étape de représentation par sacs de mots, et une étape de classification. Nous détaillons dans la suite ces différentes étapes.

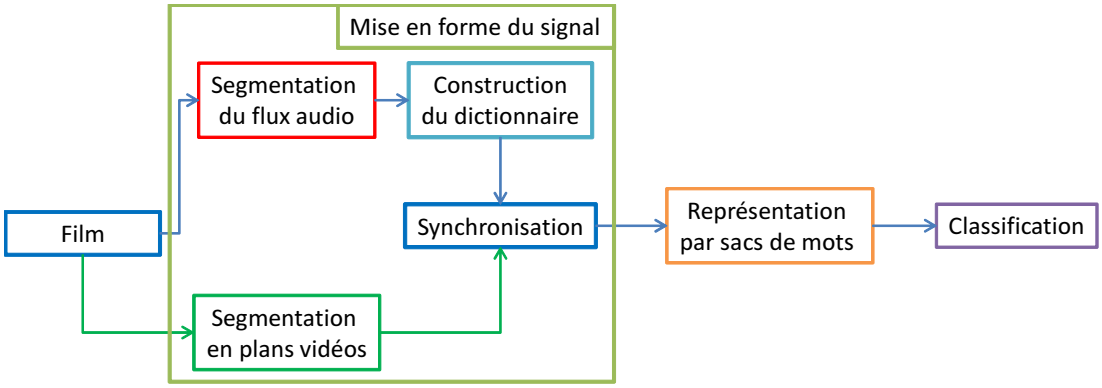


FIGURE 6.1 – Description du système mis en place pour la représentation TF-IDF. Les flèches vertes correspondent au trajet de la vidéo, et les flèches bleues au trajet de l'audio.

6.1.2.1 Mise en forme du signal

L'objectif de cette étape est de mettre en forme le signal pour pouvoir le représenter par sacs de mots.

Tout d'abord, le signal audio est segmenté, puis une quantification en mots audio est réalisée. Ces deux étapes sont identiques à ce qui est présenté dans la figure 3.1 de la section 3.2.2. Nous utilisons pour cette étape uniquement les MFCC, un seul dictionnaire, et 128 mots dans ce dictionnaire, i.e., $N = 1$, $K = 1$ et $C = 128$.

Parallèlement, une segmentation en plans de la vidéo est effectuée à l'aide du logiciel de segmentation en plans de Technicolor, fondé sur la comparaison d'histogrammes couleur entre des images successives.

L'étape de synchronisation permet ensuite d'indiquer à quel plan vidéo appartiennent les segments audio. Un segment audio appartient à un plan vidéo s'il intersecte ce dernier. Au final, par rapport au système décrit dans la section 3.2.2, nous indiquons comme information supplémentaire à chaque segment le plan vidéo auquel il appartient.

6.1.2.2 Représentation par sacs de mots

Dans cette étape, chaque plan vidéo de chaque film de la base de données MediaEval est représenté par un vecteur de poids TF-IDF. Nous calculons l'IDF sur la base des 26 108 plans vidéos de la base de données, c'est-à-dire que par analogie avec le traitement des documents textuels, nous considérons que nos 26 108 plans vidéos constituent un ensemble de 26 108 documents. Étant donnée la longueur des plans vidéos, les vecteurs obtenus sont très creux.

6.1.2.3 Classification

Pour construire un modèle des plans violents, nous utilisons des SVMs, et nous comparons deux types de noyaux particulièrement utilisés pour la comparaison d'his-

togrammes, comme dans Schiele *et al.* [117] par exemple : le noyau intersection d'histogramme (HIK), et le noyau χ^2. Si l'on définit $x, y \in \mathbb{R}^D$ comme étant des vecteurs de dimension D, alors ces deux noyaux peuvent se définir par :

$$k_{HIK}(x,y) = \sum_{i=1}^{D} \min(x_i, y_i) \qquad k_{\chi^2}(x,y) = \frac{1}{2} \sum_{i=1}^{D} \frac{x_i * y_i}{x_i + y_i}$$

Pour pouvoir calculer le MAP@100, qui est la métrique officielle, nous avons besoin de pouvoir ordonner le résultat rendu par l'algorithme, à l'aide par exemple d'une confiance sur le résultat. Pour cela, nous récupérons pour chaque échantillon x sa distance à l'hyperplan défini par le modèle. La distance à l'hyperplan est positive si le modèle classe l'échantillon comme non-violent, négative s'il le classe comme violent. De plus, si la valeur rendue est inférieure à 1, en valeur absolue, alors l'échantillon se situe à l'intérieur de la marge. La distance à l'hyperplan est ensuite normalisée entre 0 et 1 à l'aide d'une fonction sigmoïde, définie de telle sorte que $f(x) = 0.9$ pour $x = -1$ et $f(x) = 0.1$ pour $x = 1$, x étant la distance à l'hyperplan et $f(x)$ la fonction sigmoïde :

$$f(x) = \frac{1}{1 + e^{x \ln 9}} \tag{6.2}$$

Etant donné le déséquilibre des données, nous avons testé l'influence de l'insertion du biais sur le paramètre C, comme défini section 3.1, en utilisant un poids $w = 10$ sur la classe violence.

L'utilisation des SVMs pour ce problème est facilitée, d'une part, par l'utilisation de noyaux non-paramétriques associés à des histogrammes creux, et, d'autre part, par la quantité relativement faible de plans vidéos. Ces deux conditions nous permettent d'utiliser la totalité des plans de la base d'apprentissage pour apprendre notre modèle.

6.1.3 Expériences

Nous présentons dans cette section les expériences que nous avons menées et les résultats obtenus lors de la campagne MediaEval 2012.

6.1.3.1 Étude sur les paramètres

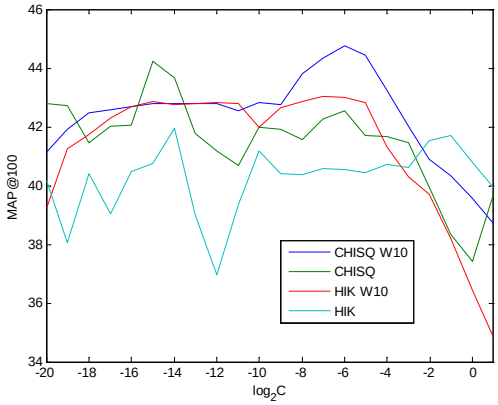
Nous étudions l'influence des différents paramètres du système par validation croisée de type $\text{CV}_{\textbf{LOMO}}$. Les valeurs de la grille de recherche du paramètre C sont fixées à intervalle constant sur une échelle logarithme, c'est-à-dire $\log_2 C \in [-20; 1], log_2 C \in \mathbb{Z}$. Nous comparons les 4 configurations suivantes :

CHISQ W10 : Utilisation du noyau χ^2 et biais $w = 10$ pour la classe violente.

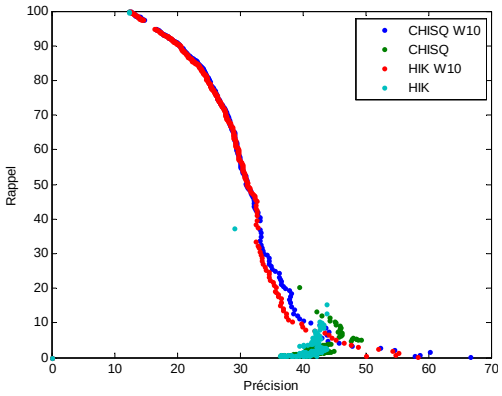
CHISQ : Utilisation du noyau χ^2 sans biais.

HIK W10 : Utilisation du noyau HIK et biais $w = 10$ pour la classe violente.

HIK : Utilisation du noyau HIK sans biais.



(a) MAP@100 obtenu sur la grille de recherche de C.



(b) Courbes rappel-précision obtenues pour $\log_2 C = -6$. Les courbes CHISQ W10 et HIK W10 se superposent pour des taux de rappels élevés.

FIGURE 6.2 – Résultats obtenus en validation croisée.

Les figures 6.2a et 6.2b montrent le MAP@100 obtenu sur la grille de recherche, et les courbes rappel précision pour $C = -6$, correspondant au MAP@100 maximum obtenu. On remarque que les modèles biaisés semble donner de bien meilleurs résultats que dans le cadre de la détection d'évènements audio. Les courbes rappel-précision des modèles non-biaisés semblent concentrées autour de 40 % de précision pour un rappel < 10 %. Ces résultats peuvent indiquer que l'action du biais est d'augmenter l'efficacité du classifieur pour des valeurs de C faibles en termes de précision et rappel. En effet, malgré ce résultat, les taux de MAP@100 obtenus pour les modèles biaisés ne sont que très légèrement supérieurs à ceux des modèles non biaisés. Il semble en revanche que la différence de performance entre le noyau χ^2 et le noyau HIK soit ténue, même si le premier semble légèrement plus performant.

La métrique officielle étant le MAP@100, nous choisissons le modèle qui nous paraît être le meilleur compromis entre valeur de MAP et déviation standard. Le modèle SVM à noyau χ^2, avec un biais $w = 10$ sur la classe violence, et $\log_2 C = -6$ est donc choisi pour être utilisé sur les films de test. Le MAP@100 en validation croisée est alors de 44,77 % pour le modèle choisi, avec une déviation standard de 15,01 %.

6.1.3.2 Résultats sur les films de test

Les résultats obtenus sont présentés dans le tableau 6.1. La colonne MAP@100 représente en fait la Précision Moyenne (AP, de l'anglais *Average Precision*) pour chacun des films, le MAP@100 étant la moyenne sur les valeurs d'AP de chaque film, présentée dans la ligne **Total**.

La première chose que l'on remarque est, encore une fois, la variabilité des résultats film par film. On remarque que l'AP@100 du film *"Dead Poet Society"* est fortement inférieure à celle des deux autres films. Ce phénomène est encore plus remarquable

Film	**P**	**R**	**F1**	**MAP@100**	**MC**
Dead Poet Society	4,79	61,76	8,90	10,85	4,09
Fight Club	26,15	65,81	37,43	52,98	3,70
Independence Day	32,87	89,49	48,08	57,77	1,35
Total	**25,00**	**77,90**	**37,85**	**40,54**	**2,50**

TABLEAU 6.1 – Résultats obtenus sur les films de test pour le système basé sur la représentation TF-IDF. La colonne P correspond à la Précision, R au Rappel, F1 à la mesure F1 et MC au MediaEval Cost.

lorsqu'on observe le rappel et la précision, ou encore le MediaEval Cost. Cela peut être dû à la variabilité dans les attributs, comme c'est le cas pour la détection de concepts, et à la variabilité dans les types de violence présentes dans les films de test. En effet, en y regardant de plus près, le film qui obtient les moins bons résultats, *"Dead Poet Society"*, est aussi celui le moins violent. Le film le plus proche de celui-ci en terme de violence dans la base d'apprentissage est probablement *"Billy Elliot"*, contenant lui aussi très peu de scènes violentes. Leur sous-représentation dans la base d'apprentissage peut expliquer les faibles résultats. Par ailleurs, les meilleurs résultats du film *"Independence Day"*, comparé à ceux du film *"Fight Club"*, peuvent s'expliquer par le fait que ce film contient plus ou moins le même type de violence que celle contenue dans le film *"Armageddon"*, à savoir des scènes contenant beaucoup d'explosions, et des scènes de destruction massive.

Malgré tout, les résultats obtenus sont très corrects surtout si l'on tient compte du fait que seul le canal audio est utilisé. Ce système très simple valide l'intérêt de la représentation des plans par des vecteurs de poids TF-IDF dans le cadre de la détection de violence.

6.1.4 Conclusions

Cette section présente un système simple basé sur la représentation en sacs de mots TF-IDF pour la détection de plans violents. Le système que nous avons mis en place permet de valider l'intérêt des représentations par sacs de mots développées dans le cadre du traitement de collection de textes. Cependant, les résultats obtenus montrent qu'il y a aussi une variabilité inter-films dans le cadre de la détection de violence, qui peut être due à la divergence statistique entre les attributs des films, ou à la subjectivité et la variabilité de la notion que l'on souhaite détecter. Cette représentation permet aussi de valider le fait que les mots audio peuvent être utilisés directement en tant que concepts dans un système de classification.

Si les résultats obtenus sont prometteurs, les représentations par sacs de mots ne permettent pas d'obtenir une segmentation fine des films, car elles nécessitent d'être calculées sur des segments temporels conséquents pour que les histogrammes extraits soient significatifs.

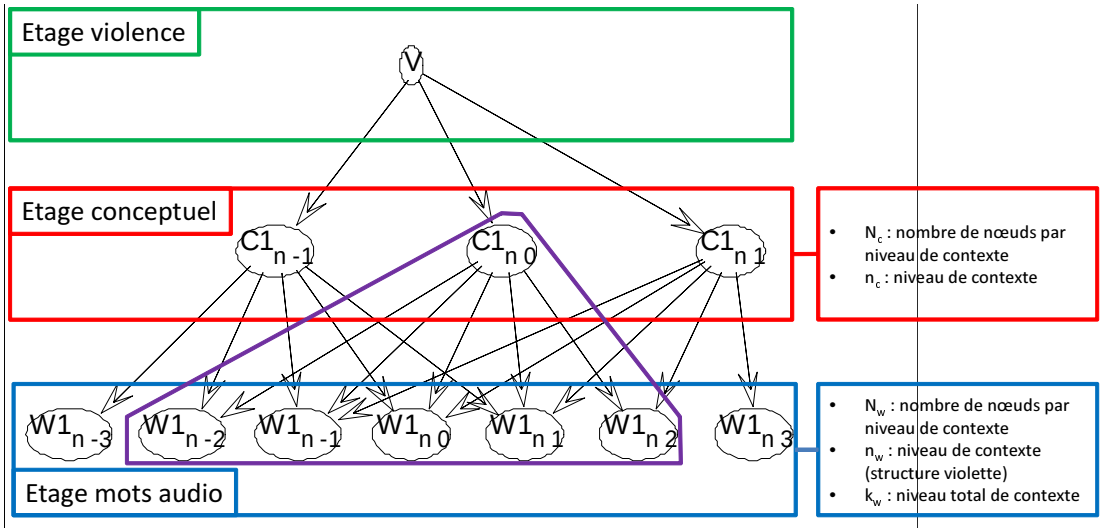


FIGURE 6.3 – Exemple de graphe contextuel hiérarchique naïf et variables utilisées pour le définir.

6.2 Utilisation de détecteurs de concepts

Dans cette section, nos motivations sont multiples. Nous souhaitons tout d'abord appliquer les conclusions de l'équipe ARF [118, 61] aux réseaux bayésiens, c'est-à-dire que nous souhaitons vérifier que des détecteurs de concepts supervisés même imparfaits peuvent quand même être utiles pour détecter la violence dans les films. Nous souhaitons aussi intégrer l'information temporelle du signal audio dans le modèle et profiter de la représentation par mots audio introduite dans la partie précédente. Enfin, nous souhaitons être capables de détecter directement des segments violents, c'est-à-dire que nous ne souhaitons pas être dépendants de la segmentation en plans vidéos, contrairement au système précédent.

Nous proposons donc dans cette section une modification du système présenté dans la section 3.2 dans le but de détecter la violence. Nous commençons par présenter la modification proposée, puis nous présentons les résultats obtenus.

6.2.1 Présentation du système

La modification que nous avons introduite dans le système de la section 3.2 porte sur l'étape de classification. Nous nous sommes inspirés de l'équipe ARF qui utilise les sorties de leurs détecteurs de concepts comme attributs de leur détecteur de violence. Nous avons tenté de reproduire cette idée en utilisant un réseau bayésien contextuel hiérarchique naïf à trois étages. Ce type de structure a l'avantage de permettre d'intégrer du contexte à la fois au niveau de la séquence de mots audio et au niveau de la séquence de concepts. Un exemple de la structure mise place est présenté sur la figure 6.3.

Le premier étage de ce graphe est l'étage *mots audio*, à la base du graphe, représenté par les nœuds $\mathrm{W1}_{\mathrm{n}\ k_w}$, $k_w \in [-3, 3]$. Cet étage représente la séquence de mots audio

extraits du signal audio. A chaque niveau temporel, nous extrayons K mots audio par dictionnaire, et nous créons N dictionnaires, c'est-à-dire que KN mots sont extraits pour chaque segments, soit KN nœuds. Le niveau de contexte à cet étage est défini par n_w, correspondant au niveau de contexte de la structure entourée en violet sur la figure 6.3. Dans le cadre des expériences menées, nous avons utilisé les MFCC avec $N = 3$ et $K = 1$.

Au milieu du graphe se trouve l'étage *conceptuel*, représenté par les nœuds $\mathrm{C1}_{\mathsf{n}\ n_c}$, $n_c \in [-1, 1]$, correspondant à la séquence de concepts coups de feu, explosions ou autres. Nous avons introduit deux types de niveau temporel : soit un seul nœud ($N_c = 1$) pour représenter tous les concepts, comme dans le système original décrit section 3.2, soit un nœud binaire par concept ($N_c = 3$ dans notre application) représentant pour chaque segment audio la présence ou non d'un des concepts. L'avantage de la deuxième représentation est qu'elle permet d'associer plusieurs évènements à un même segment. Ainsi, si un segment audio contient la réalisation d'une explosion en même temps que d'un coup de feu, il est ainsi possible de le modéliser. En revanche, les besoins en mémoire sont beaucoup plus importants, et on ne peut pas dépasser un niveau de contexte supérieur à $n_c = 1$ sur l'étage conceptuel (pour plus d'information, voir annexe B). Les nœuds $\mathrm{C1}_{\mathsf{n}\ n_c}$ concepts à cet étage sont reliés de manière naïve aux nœuds $\mathrm{W1}_{\mathsf{n}\ k_w}$, $k_w \in [-n_c - n_w; n_c + n_w]$. Par exemple, sur la figure 6.3, le nœud $\mathrm{C1}_{\mathsf{n}\ 0}$ est relié aux nœuds $\mathrm{W1}_{\mathsf{n}\ -2}$ à $\mathrm{W1}_{\mathsf{n}\ 2}$.

Enfin, le dernier étage est l'étage `violence`, tout en haut du graphe, représenté par le nœud V. Ce nœud permet de déterminer pour chaque segment s'il est violent ou non violent. Il est relié à tous les nœuds concepts de l'étage conceptuel à la manière d'un réseau bayésien naïf classique.

Si cette structure nous permet d'implémenter une idée semblable à l'idée d'ARF mais fondée sur des réseaux bayésiens plutôt que sur des réseaux de neurones, leur systeme utilisait les sorties des détecteurs de concepts comme attributs en entrée du détecteur de violence, à la fois pendant la phase d'apprentissage et pendant la phase de test. Dans notre système, nous utilisons directement les annotations du jeu de données ME-A pour apprendre les paramètres à tous les étages du système, car utiliser les sorties du détecteur de concepts nécessiterait d'utiliser des variables continues, et nous n'avons pas exploré ce champ des réseaux bayésiens. En revanche, lors de la phase de test, seules les valeurs des mots audio sont observées, et la phase d'inférence doit donc déterminer les valeurs des nœuds concepts avant de déterminer les valeurs du nœud violence. Ainsi, l'idée d'ARF est de fait implémentée dans la phase d'inférence mais pas dans la phase d'apprentissage.

Pour comparer les résultats obtenus avec ceux proposés par ARF, nous procédons ensuite à la même agrégation en plans vidéos : la probabilité qu'un plan soit violent est la probabilité maximale obtenue sur les segments appartenant à ce plan, ce qui correspond à une hypothèse optimiste d'agrégation.

6.2.2 Expériences

Nous présentons dans cette section les expériences réalisées et les résultats obtenus, tant au niveau de la détection de violence qu'au niveau de la détection concepts. Pour des raisons d'utilisation de mémoire vive, nous nous sommes limités à une profondeur de contexte de $N_c \leq 1$ (voir annexe B pour plus d'explications) :

#1. $\underline{n_c = 0, N_c = 3}$: pas de contexte à l'étage conceptuel, et trois nœuds binaires utilisés à cet étage.

#2. $\underline{n_c = 0, N_c = 1}$: pas de contexte à l'étage conceptuel, et un nœud multiclasse utilisé à cet étage.

#3. $\underline{n_c = 1, N_c = 3}$: niveau de contexte compris entre $n_c - 1$ et $n_c + 1$, et trois nœuds binaires utilisés à cet étage.

#4. $\underline{n_c = 1, N_c = 1}$: niveau de contexte compris entre $n_c - 1$ et $n_c + 1$, et un nœud multiclasse utilisé à cet étage.

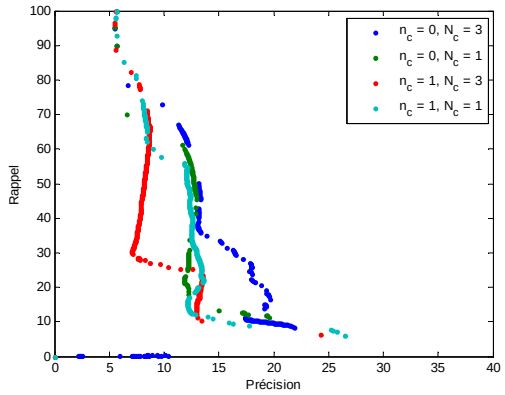
Les autres paramètres sont fixés par défaut à l'utilisation des MFCC, $N = 3$, $K = 1$ et $n_w = 5$.

Nous présentons tout d'abord les résultats obtenus par validation croisée de type $\text{CV}_{\textbf{LOMO}}$, avant de présenter les résultats obtenus sur les films de test.

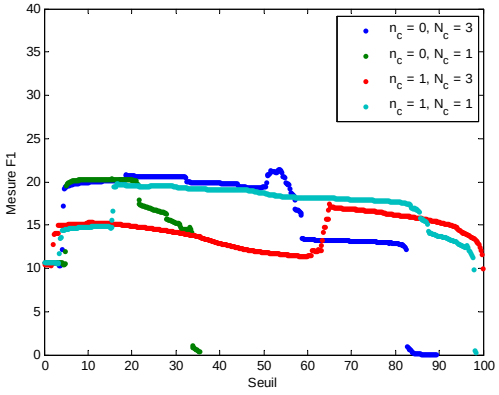
6.2.2.1 Validation croisée

Détection de violence :

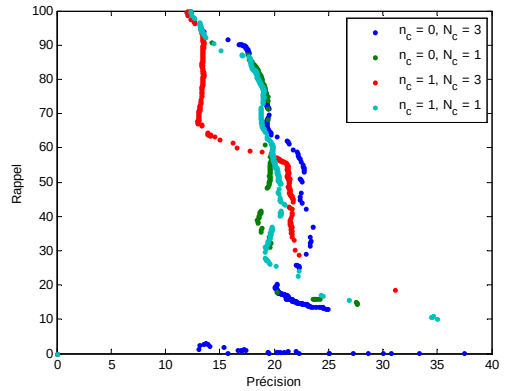
Les résultats obtenus sont présentés sur les figures 6.4a et 6.4b pour le niveau segments, et sur les figures 6.4c et 6.4d pour le niveau plans vidéos. Les courbes de rappel-précision ainsi que les courbes de mesure F1 sont reportées pour chaque expérience. Des résultats quantitatifs sont aussi présentés dans le tableau 6.4e. On remarque tout d'abord que les trois systèmes ont des comportements différents. Le système #1 semble être celui qui donne les meilleures performances en terme de précision, tandis le système #3, qui ajoute un niveau de contexte au système #1 à l'étage conceptuel, semble être le moins performant pour des taux de rappel compris entre $\simeq$60 % et $\simeq$25 %. Nous n'expliquons pas la raison de ce comportement. Les systèmes #2 et #4 semblent avoir des performances sensiblement équivalentes en terme de rappel-précision, mais les courbes de mesure F1 montrent clairement que le système #4 est bien plus robuste au seuil de probabilité que le système #2. Cela démontre encore une fois l'intérêt de la prise en compte de l'évolution temporelle pour prendre une décision. Les performances du système #4 sont sensiblement équivalentes à celles du système #1 pour des seuils de probabilité assez bas, légèrement meilleures pour des seuils plus élevés. Il semble de plus que l'agrégation en plans améliore les résultats obtenus : la mesure F1 maximum augmente d'environ 10 % pour chacune des expériences, ce qui signifie un rappel et une précision améliorés. Cela reste logique car l'hypothèse d'agrégation faite est une hypothèse optimiste. En revanche, il semble que ce système soit moins performant que la représentation vectorielle comme l'indiquent les taux de MAP@100 environ deux fois inférieurs ainsi que les moins bonnes courbes rappel-précision.



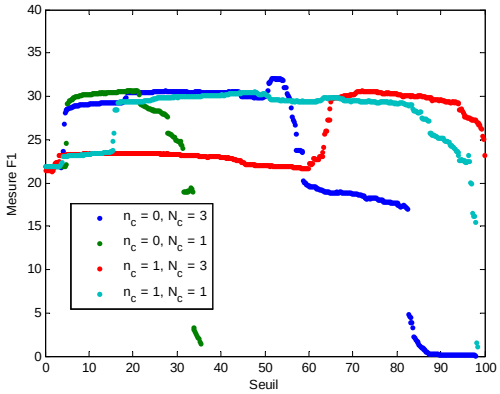
(a) Précision vs rappel, niveau segments audio.



(b) Mesure F1, niveau segments audio.



(c) Précision vs rappel, niveau plans.



(d) Mesure F1, niveau plans.

	SEGMENTS				**SHOTS**			
	#1	#2	#3	#4	#1	#2	#3	#4
MAP@100	11,22	**15,39**	14,64	12,28	19,36	15,66	**22,08**	18,02
STD	**11,02**	24,81	16,75	23,81	13,89	14,90	**13,37**	17,72
Mesure F1	**21,46**	20,35	17,43	19,81	**32,15**	30,75	30,64	30,53
Précision	**17,80**	12,66	13,37	12,08	**22,11**	19,36	21,18	19,05
Rappel	27,01	51,78	25,04	**54,88**	58,89	74,86	55,46	**76,76**

(e) Résultats quantitatifs. STD correspond à la déviation standard de la précision moyenne en validation croisée.

FIGURE 6.4 – Résultats obtenus pour la détection de violence en validation croisée. Comparaison avec l'agrégation en plans.

Comparé au système précédent, il faut noter que le nombre de concepts utilisés pour représenter le flux audio est faible. La représentation conceptuelle du film est donc bien plus limitée. De plus, cela restreint la violence aux seuls concepts coups de feu et explosions, ce qui ne correspond pas forcément à toutes les scènes violentes. Cela

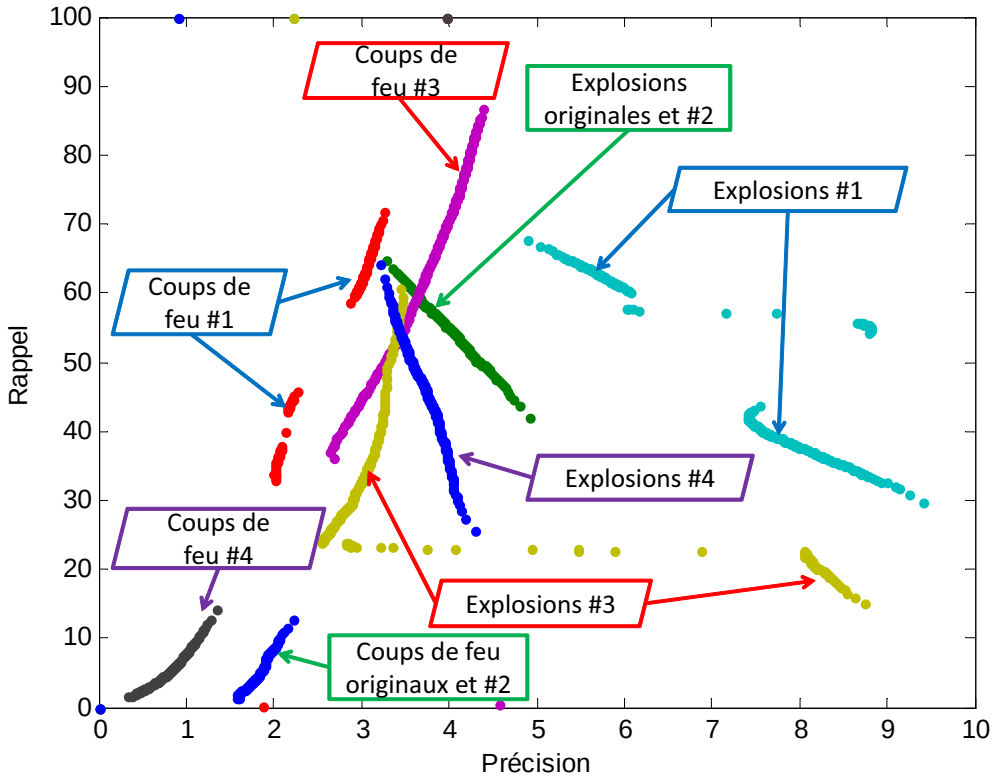


FIGURE 6.5 – Influence du système sur la détection de concepts. Les courbes de l'expérience #2 sont identiques aux courbes originales.

peut expliquer les moins bons résultats obtenus avec ce système.

Influence sur la détection de concepts :

La figure 6.5 présente les résultats obtenus pour la détection de coups de feu et d'explosions. Nous y comparons les résultats de détection de concepts des systèmes #1, #3 et #4, avec ceux obtenus avec le détecteur de concept original développé dans la partie précédente [1]. Les résultats du système #2 sont identiques aux résultats originaux, ce qui semble logique : le nœud multiclasse unique de l'étage conceptuel est seul connecté à l'étage violence. Ainsi, si on omet le dernier étage, on obtient le détecteur de concepts original. Le système #4, qui est le plus proche de notre système original en ajoutant des niveau de contexte à l'étage conceptuel, semble donner des résultats inférieurs en terme de précision par rapport à ceux obtenus avec le système original ce qui indiquerait que cela a une influence néfaste sur la qualité de la détection de coups de feu et d'explosions. En revanche, pour les expériences #1 et #3, le réseau utilisé semble avoir une grande influence sur la détection de coups de feu et d'explosions par rapport au système

1. Pour la comparaison, nous avons recalculé les résultats obtenus précédemment, en effectuant une validation croisée sur les 15 films du jeu de données ME-A. La différence de résultats s'explique par la différence obtenue précédemment entre les films de test et ceux d'apprentissage.

original à un nœud. Dans l'ensemble, le système #1 semble privilégier les explosions en terme de précision, tandis que le système #3 semble privilégier le rappel des coups de feu au détriment de la précision des explosions. Cela peut indiquer que le contexte au niveau des concepts binaires profite aux coups de feu, au contraire des explosions. Cela peut aussi indiquer que la relation entre deux segments audio consécutifs de type explosions n'est pas contenue dans les mots audio extraits des MFCC. Il semble aussi que les résultats des systèmes #1 et #3 soient beaucoup moins robustes par rapport au seuil de probabilité, mais nous n'avons pas d'explications quant à ce phénomène.

6.2.2.2 Résultats sur les films de test

Les résultats obtenus sur les films de test sont présentés sur les figures 6.6a et 6.6b pour le niveau segments, et sur les figures 6.6c et 6.6d pour le niveau plans vidéos. Des résultats quantitatifs sont aussi présentés dans le tableau 6.6e. Les résultats sont sensiblement équivalents à ceux obtenus en validation croisée pour les systèmes #2, #3 et #4. Le système #1 est en revanche moins performant, en particulier pour des seuils de probabilité élevés. Le système #4 obtient le meilleur MAP@100 sur les films de test, à la fois au niveau plans et au niveau segments, et les mauvais résultats du système #2, tant au niveau segments qu'au niveau plans semblent aussi confirmer l'intérêt de l'intégration temporelle pour prendre une décision.

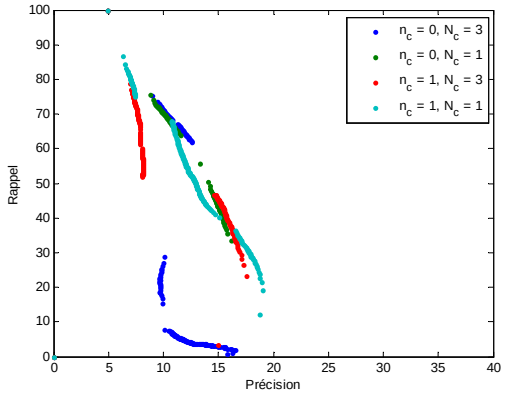
Comme indiqué dans la section précédente, les moins bons résultats de notre système en terme de MAP@100 par rapport à la représentation vectorielle peuvent s'expliquer par une représentation peut-être trop stricte à l'étage conceptuel. En effet, la représentation vectorielle peut s'apparenter à utiliser C "concepts" pour représenter la violence. Notre système réduit ces C concepts à trois : autres, coups de feu et explosions, ces derniers n'étant pas forcément représentatifs de toutes les scènes violentes.

6.2.3 Conclusions

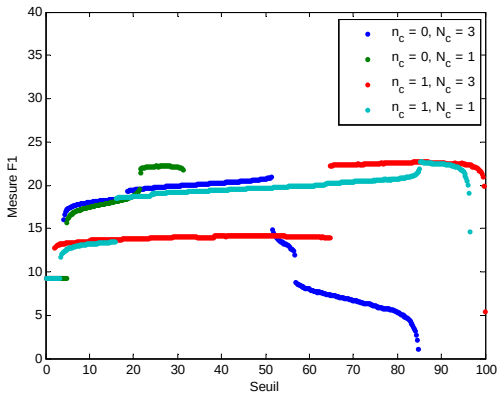
Cette section présente un système de détection de scènes violentes basé sur l'utilisation de détecteurs de concepts et sur l'utilisation de réseaux bayésiens contextuels hiérarchiques. Ce système nous permet de vérifier que nous pouvons obtenir un détecteur de violence correct fondé sur des détecteurs de concepts relativement faibles. Le système que nous proposons a de plus l'avantage de fournir des décisions pour chaque segment audio.

Les résultats que nous obtenons sont corrects en terme de rappel et de précision en restant inférieurs à ceux obtenus par représentation vectorielle, mais les valeurs de MAP@100 obtenues sont nettement inférieures. Ce résultat peut s'expliquer par le trop faible nombre de concepts utilisés dans notre système.

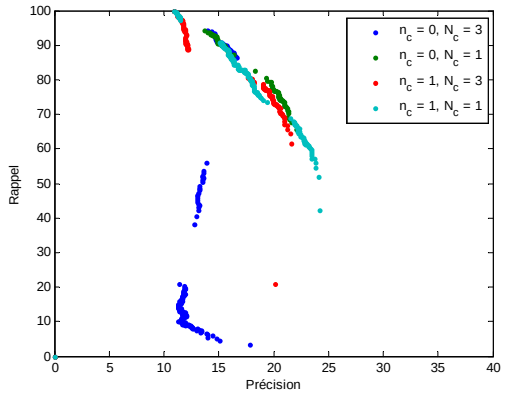
Malgré des résultats moins bons que ceux que nous attendions, ces systèmes semblent quand même encourageants, et semblent aller dans le sens des affirmations d'ARF : il paraît ainsi possible d'obtenir un détecteur de scènes violentes correct fondé sur de faibles détecteurs de concepts audio.



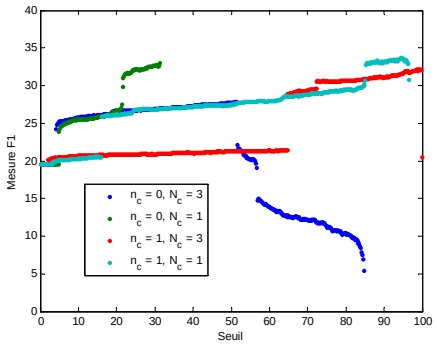
(a) Précision vs rappel, niveau segment audio.



(b) Mesure F1, niveau segment audio.



(c) Précision vs rappel, niveau plan.



(d) Mesure F1, niveau plan.

	SEGMENTS				**SHOTS**			
	#1	#2	#3	#4	#1	#2	#3	#4
MAP@100	12,15	7,34	**14,20**	11,72	14,41	12,10	**20,84**	15,83
STDAP@100	18,94	**8,56**	13,68	10,33	16,16	11,89	**9,36**	15,48
Mesure F1	20,95	22,27	22,72	**22,80**	27,94	32,99	32,26	**33,71**
Précision	12,61	15,02	15,45	**16,58**	16,67	22,04	21,50	**23,47**
Rappel	**61,85**	43,11	42,91	36,49	**86,29**	43,11	64,62	59,86

(e) Résultats quantitatifs. STD correspond à la déviation standard de la précision moyenne sur les films de test.

FIGURE 6.6 – Résultats obtenus pour la détection de violence sur les films de test. Comparaison avec l'agrégation en plans.

6.3 Apprentissage de structure et multimodalité

Nous proposons dans cette section un système multimodal de détection de violence, c'est-à-dire un système fondé sur l'utilisation de l'audio et de la vidéo. Ce système,

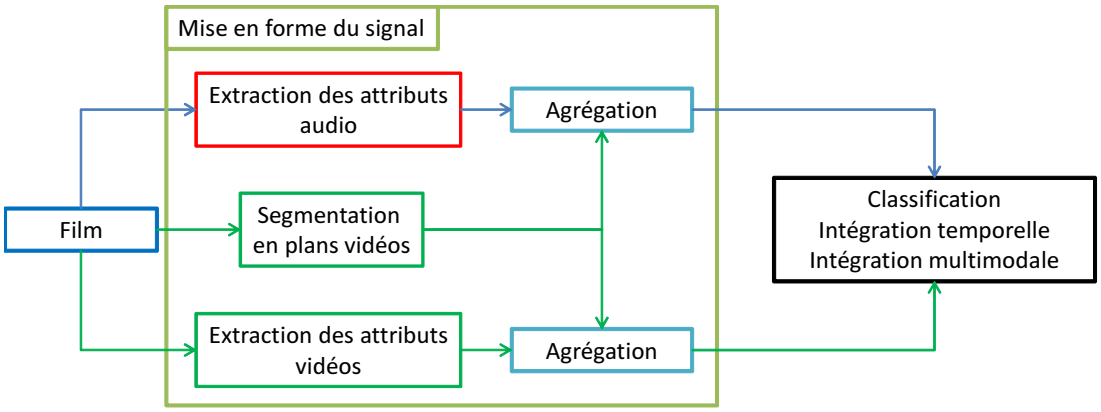


FIGURE 6.7 – Présentation du système basé sur l'apprentissage de structure.

développé dans le cadre de MediaEval 2011, a conduit à une publication à ICASSP en 2012 [106]. Ce système a été reconduit pour MediaEval 2012 moyennant quelques modifications légères, notamment dans le choix des attributs. Il est fondé sur la détection de plans violents.

Nous évaluons avec ce système l'influence de l'apprentissage de structure dans les réseaux bayésiens, pour tenter de confirmer les travaux de Baghdadi [5, 57]. Nous étudions aussi l'influence de l'intégration temporelle au travers des réseaux bayésiens contextuels et du filtrage temporel, et l'intégration multimodale en comparant fusions précoce et tardive.

Nous présentons dans cette section la version 2012 du système ainsi que les résultats obtenus.

6.3.1 Présentation du système

Le système est présenté sur la figure 6.7 et se décompose en 4 parties : la mise en forme du signal, la classification, l'intégration temporelle et l'intégration multimodale. Concernant les trois dernières parties, nous avons testé toutes les combinaisons de classification, avec ou sans intégration temporelle précoce et/ou tardive, avec ou sans intégration multimodale précoce et/ou tardive.

6.3.1.1 Mise en forme du signal

Nous avons choisi d'extraire différents types d'attributs bas niveau du signal audio et du signal vidéo, et ce, pour chaque plan. Les attributs que nous avons choisis sont des attributs bas niveau pour lesquels il est aisé d'obtenir une valeur par plan en utilisant par exemple l'agrégation temporelle à l'aide de moyennes sur la durée du plan.

Attributs audio : Les attributs audio, extraits de fenêtres de 40 ms avec 20 ms de recouvrement, sont : l'énergie (E), le barycentre (C), l'asymétrie (A), la platitude (F) et la fréquence de coupure à 90 % (R) du spectre, et le taux de passage par zéro (Z) du signal. Ces attributs sont ensuite centrés réduits et moyennés sur

Attributs audio		**Attributs vidéos**	
Energie	E	Durée des plans	SL
Barycentre	C	Proportion de sang	B
Asymétrie	A	Activité moyenne	AC
Platitude	F	Nombre de flashs	FL
Fréquence de coupure	R	Proportion de feu	FI
Taux de passages par zéro	Z	Cohérence couleur	CC
		Patron majoritaire	Tp
		Angle moyen	Al
		Energie du patron	Em

TABLEAU 6.2 – Récapitulatif des attributs audio et vidéos utilisés, et de leurs acronymes.

la durée des plans vidéos, de façon a obtenir une valeur par plan. Le vecteur d'attributs ainsi obtenu à une dimension $D = 6$.

Attributs vidéos : Les attributs vidéos extraits sont, pour chaque plan vidéo, la durée du plan (SL), la proportion moyenne de pixels de couleur sang (B), l'activité moyenne (AC), le nombre de flashs (FL), la proportion moyenne de pixels de couleur feu (FI), une mesure de cohérence de couleur (CC), la luminance moyenne (AVL) et trois attributs d'harmonie couleur, à savoir le patron harmonieux majoritaire (Tp), l'angle moyen du patron majoritaire (Al), et l'énergie du patron majoritaire (Em) [6]. Le vecteur d'attributs ainsi obtenu a une dimension $D = 10$.

Les valeurs des attributs sont ensuite discrétisées film par film (sauf pour le patron majoritaire, dont les valeurs sont déjà discrétisées sur 9 valeurs), pour obtenir 21 valeurs pour chacun des attributs. Le tableau 6.2 présente un récapitulatif des attributs audio et vidéos utilisés pour ce système.

6.3.1.2 Classification

Nous utilisons pour ce système les réseaux bayésiens, en nous focalisant sur la comparaison de la structure naïve, "faite à la main", et de structures apprises à l'aide de scores descriptifs :

- La structure FAN, introduite par Lucas [92].
- La structure résultant de l'algorithme K2 [34], nécessitant un ordonnancement des nœuds du graphe, mais étant plus générique que la structure FAN.

6.3.1.3 Intégration temporelle

L'intégration temporelle est prise en compte de deux manières différentes : par l'utilisation d'attributs contextuels, et/ou par l'utilisation d'un filtre temporel sur la sortie du classifieur. Nous comparons les résultats, avec et sans contextualité, pour un niveau de contexte de $n = 5$.

Nous avons testé les filtres temporels suivants :

– Prendre le vote majoritaire sur une fenêtre d'échantillons de taille $k = 5$, après seuillage des probabilités.
– Moyenner les probabilités sur une fenêtre glissante de taille $k = 5$, avant seuillage des probabilités.

La métrique officielle dans le cadre de MediaEval étant le MAP@100, il faut pouvoir ordonner les sorties du système, c'est-à-dire qu'il nous faut obtenir pour chaque échantillon une valeur numérique nous indiquant le degré de violence de l'échantillon estimé par le classifieur. Ainsi, dans le cas du vote majoritaire, il nous faut redéfinir cette valeur de confiance. Pour chaque fenêtre de vote majoritaire, nous considérons donc deux cas pour associer une valeur de violence aux échantillons, selon que le résultat du vote majoritaire est violent ou non violent :

Vote majoritaire violent : Les valeurs de violence des échantillons non violents à l'intérieur de la fenêtre sont positionnées à $\min(P(S_v))$, où $P(S_v)$ est l'ensemble des probabilités des échantillons violents dans la fenêtre.

Vote majoritaire non violent : Les valeurs de violence des échantillons violents à l'intérieur de la fenêtre sont positionnées à $\max(P(S_{nv}))$, où $P(S_{nv})$ est l'ensemble des probabilités des échantillons non violents dans la fenêtre.

6.3.1.4 Intégration multimodale

Pour l'intégration multimodale, nous comparons la fusion précoce et la fusion tardive. La fusion précoce consiste simplement à concaténer les vecteurs d'attributs vidéos et audio, et à effectuer l'apprentissage avec les attributs concaténés. Pour la fusion tardive, les sorties des classifieurs pour le i$^{\text{ème}}$ plan vidéo sont fusionnées à l'aide de la formule suivante :

$$P^{s_i}_{fused}(P^{s_i}_{v_a}, P^{s_i}_{v_v}) = \begin{cases} max(P^{s_i}_{v_a}, P^{s_i}_{v_v}) & \text{si les deux sont violents} \\ min(P^{s_i}_{v_a}, P^{s_i}_{v_v}) & \text{si les deux sont non violents} \\ P^{s_i}_{v_a} * P^{s_i}_{v_v} & \text{autrement} \end{cases} \tag{6.3}$$

où $P^{s_i}_{v_a}$ (respectivement $P^{s_i}_{v_v}$) est la probabilité que le plan i soit violent par rapport au classifieur audio (respectivement vidéo). Cette règle donne un score élevé au plan quand les deux classifieurs indiquent que le plan est violent, un score faible quand ceux-ci indiquent qu'il n'est pas violent, et enfin un score intermédiaire s'ils prennent des décisions différentes.

6.3.2 Expériences

Dans cette section nous présentons les résultats que nous avons obtenus en validation croisée, puis nous présentons la combinaison de paramètres choisie et les résultats obtenus sur les films de test. Enfin, nous présentons et expliquons les réseaux obtenus.

6.3.2.1 Résultats obtenus

Dans cette section nous présentons les résultats obtenus par validation croisée de type $\text{CV}_{\textbf{LOMO}}$. Nous présentons les valeurs de MAP@100 obtenues dans le tableau 6.3.

Struct.	**Cont.**	**Audio**			**Vidéo**			**Fusion précoce**		
		1	2	3	1	2	3	1	2	3
Naïve	Non	36,3	**39,4**	38,4	25,4	**30,0**	27,9	36,0	**40,3**	37,5
	Oui	36,9	36,2	**37,3**	31,1	30,8	**31,3**	38,5	37,1	**38,5**
FAN	Non	26,9	**30,9**	29,3	22,4	**26,9**	25,0	29,0	34,7	**34,8**
	Oui	20,1	20,6	**21,4**	25,5	**27,4**	26,9	25,6	**26,2**	26,1
K2	Non	36,3	**39,1**	37,8	26,0	**30,7**	29,0	37,4	**40,9**	39,2
	Oui	36,1	**39,0**	37,0	27,0	27,5	**27,9**	32,3	32,3	**33,2**

TABLEAU 6.3 – MAP@100 obtenus par validation croisée pour le système multimodal. Les résultats sont rapportés pour les modalités audio et vidéo et la fusion précoce. Pour chaque modalité, la colonne 1 correspond à une absence de filtre temporel, la colonne 2 à une moyenne sur fenêtre glissante, et la colonne 3 à un vote majoritaire. Struct. : Structure utilisée, Cont. : Contexte.

$\mathbf{S_a}$	$\mathbf{C_a}$	$\mathbf{S_v}$	$\mathbf{C_v}$	$\mathbf{T_c}$	$\mathbf{T_{lf}}$	**MAP@100**
K2	Non	Naïve	Oui	1	2	43,18
K2	Oui	Naïve	Oui	3	2	42,59
K2	Oui	Naïve	Oui	1	2	42,55
K2	Oui	Naïve	Oui	2	2	42,53
Naïve	Non	Naïve	Oui	3	2	42,45
K2	Non	Naïve	Oui	3	3	42,36
Naïve	Non	Naïve	Oui	3	3	42,32

TABLEAU 6.4 – Résultats obtenus pour la fusion tardive, pour les 7 meilleures combinaisons de paramètres. $\mathbf{S_a}$: Structure audio, $\mathbf{C_a}$: Contexte audio, $\mathbf{S_v}$: Structure vidéo, $\mathbf{C_v}$: Contexte vidéo, $\mathbf{T_c}$: filtre temporel appliqué aux classifieurs, $\mathbf{T_{lf}}$: filtre temporel après fusion.

Pour la fusion tardive, toutes les combinaisons de classifieurs possibles, c'est-à-dire structure naïve, FAN ou K2, avec ou sans contexte, avec ou sans filtre temporel [2], ont été testées et les sept meilleures combinaisons sont présentées dans le tableau 6.4.

Il est intéressant de noter dans un premier temps que les réseaux FAN, bien que censés être plus performants en classification, sont surclassés par les réseaux K2 et naïf, quels que soient la modalité ou le filtre temporel considérés comme pour le système de détection d'évènements présenté section 3.2. Les structures naïves et K2 semblent quant à elles sensiblement équivalentes en termes de résultats, montrant que l'apprentissage de structure n'est pas forcément bénéfique en termes de résultats. Côté temporel, si l'importance du contexte n'est pas toujours claire pour les modalités présentes dans le tableau 6.3, les filtres temporels améliorent systématiquement les résultats, ce qui montre bien l'importance de prendre en compte l'aspect temporel du signal. Il n'est en

2. Le même filtre est appliqué aux deux classifieurs.

Film	**P**	**R**	**F1**	**MAP@100**	**MC**
Dead Poet Society	5,06	64,71	9,38	60,56	4,09
Fight Club	25,14	58,06	35,09	53,15	3,70
Independence Day	26,22	75,20	38,89	71,76	1,35
Total	**21,72**	**67,27**	**32,83**	**61,82**	**3,57**

TABLEAU 6.5 – Résultats obtenus sur les films de test pour le système multimodal. La colonne P correspond à la Précision, R au Rappel, F1 à la mesure F1 et MC au MediaEval Cost.

revanche pas possible de dire quel filtre est plus performant que l'autre. Enfin, l'importance de l'intégration multimodale est montrée par les meilleurs résultats obtenus presque systématiquement par les fusions tardive et précoce. La fusion tardive renforce de plus l'importance de l'intégration temporelle : parmi les sept meilleures expériences, le réseau naïf contextuel est utilisé pour la vidéo et un filtre temporel est systématiquement appliqué après la fusion. Il semble de plus que la fusion tardive soit plus performante que la fusion précoce.

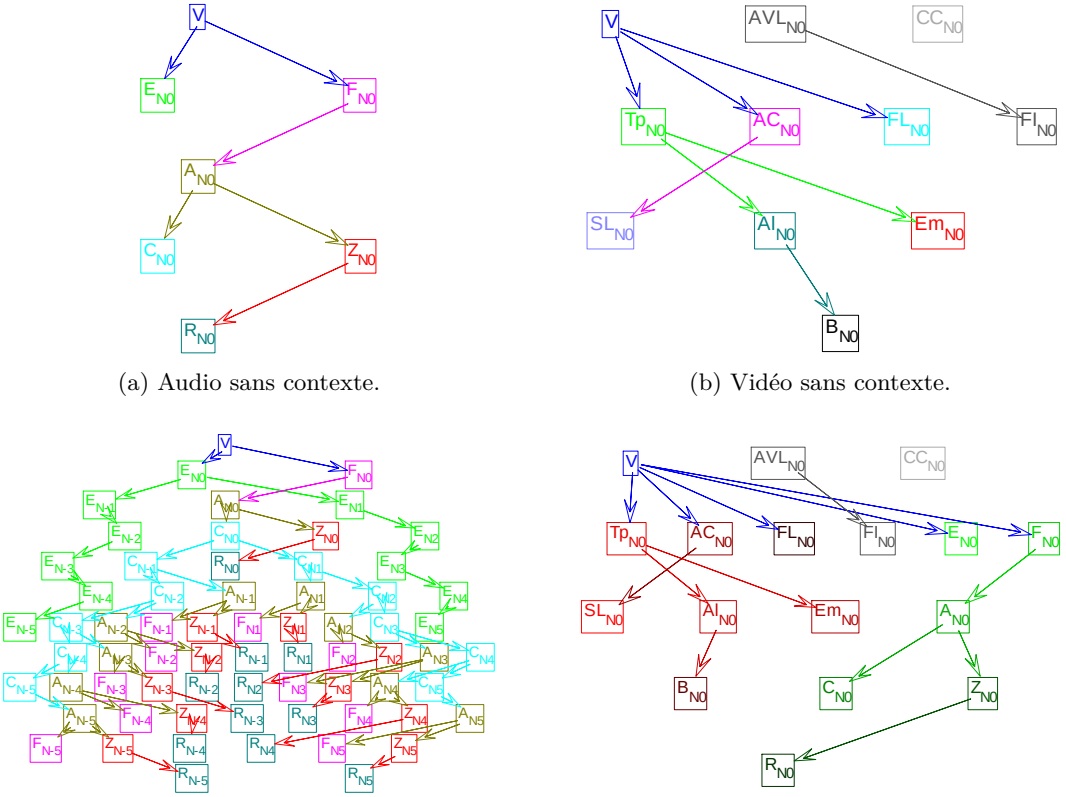
Au final, le système choisi et soumis pour la campagne 2012 est le meilleur système obtenu par fusion tardive, utilisant un réseau K2 non contextuel pour l'audio, un réseau naïf contextuel pour la vidéo, et un filtre moyennant la probabilité après fusion. Ce système a été appliqué aux films de test, et les résultats sont présentés dans le tableau 6.5. La première chose à noter est que le MAP@100 obtenu est bien meilleur que l'estimation par validation croisée ($\simeq +18\,\%$). On remarque que contrairement aux systèmes décrits dans les sections précédentes, la moins bonne AP@100 est obtenue sur le film *"Fight Club"* au lieu du film *"Dead Poet Society"*. Ce dernier obtient une valeur de MAP@100 très supérieure aux précédentes valeurs obtenues, et cela malgré une valeur de mesure F1 sensiblement identique. Cela montre clairement que les mesures d'évaluation utilisées ne sont pas équivalentes, et qu'il est intéressant de diversifier les métriques d'évaluations pour juger la qualité des résultats présentés.

De même que pour le système basé sur la représentation vectorielle, les bons résultats du film *"Independence Day"* s'expliquent par sa ressemblance avec *"Armageddon"*, dans la base d'apprentissage.

Enfin, nous pensons que ces résultats montrent l'importance de l'intégration multimodale dans ce type de système. En effet, en validation croisée, l'intégration tardive surpasse la meilleure combinaison monomodale de $\simeq +3\,\%$, tandis que la fusion précoce surpasse celle-ci d'environ $\simeq +1\,\%$.

6.3.2.2 Analyse des graphes obtenus

Les résultats les plus intéressants sont peut-être ceux obtenus en observant les graphes, particulièrement ceux issus de l'apprentissage K2 parce qu'ils nous montrent des liens existants entre les attributs et entre les attributs et la violence. Les figures 6.8a, 6.8b, 6.8c et 6.8d montrent des exemples de graphes appris par algorithme K2. Il est facile d'interpréter les résultats obtenus. Par exemple, pour le graphe vidéo sans contexte,



(a) Audio sans contexte.

(b) Vidéo sans contexte.

(c) Audio avec contexte.

(d) Fusion précoce sans contexte.

FIGURE 6.8 – Exemples de graphes obtenus par apprentissage de structure de type K2. Pour les graphes audio et vidéo, une couleur correspond à un attribut. Pour la fusion précoce, une couleur correspond à une modalité.

le lien entre l'activité et la durée du plan s'explique par le fait que le détecteur de plans utilisé a tendance à sur-segmenter quand l'activité est importante. On remarque aussi que les trois attributs d'harmonie couleur sont liés entre eux, et que l'attribut sang, dépendant de la couleur, y est aussi lié. Le lien entre luminance moyenne et feu s'explique aussi par l'augmentation de la luminance en présence de feu. L'algorithme est aussi capable de reconnaître que l'ensemble (luminance moyenne + feu) et la cohérence moyenne ne sont pas dépendants des autres variables en les déconnectant du graphe principal. On remarque ensuite que l'algorithme produit une structure temporelle stricte, comme le montre le graphe audio avec contexte : les attributs au temps $t = n$ sont liés ensemble, et non à des attributs temporels différents, à l'exception des attributs formant des chaînes temporelles logiques, comme c'est le cas pour l'énergie ou le barycentre par exemple. On remarque que pour $t \neq 0$, chaque attribut barycentre est connecté à l'attribut asymétrie correspondant, lui-même connecté à la platitude d'une part, et au taux de passages par zéro, puis la fréquence de coupure, d'autre part. Le graphe obtenu par fusion précoce sans contexte est en fait une simple "concaténation"

des graphes audio et vidéo sans contexte. Cela montre sûrement qu'il n'y a pas de lien direct, ou évident, entre les attributs audio et vidéos que nous avons utilisés. Plus généralement, les graphes obtenus montrent bien la bonne capacité de description des algorithmes d'apprentissage de structure.

6.3.3 Conclusions

Nous présentons dans cette section notre meilleur système MediaEval 2012, basé sur l'apprentissage de structure et sur l'intégration temporelle et multimodale. Les résultats obtenus sont bons par rapport aux résultats précédemment obtenus et montrent l'intérêt des représentations multimodales. En effet, les graphes multimodaux surclassent les graphes monomodaux dans presque tous les cas. Nous montrons aussi que l'apprentissage de structure descriptive permet d'obtenir des graphes logiques, et faciles à interpréter. En revanche, il semble qu'en fonction des attributs du graphe, l'apprentissage de structure ne soit pas toujours bénéfique, et que parfois une simple structure naïve donne de meilleurs résultats qu'une structure apprise à l'aide des données disponibles.

6.4 Comparaison à l'état de l'art : résultats de la campagne MediaEval 2012

Dans cette section, nous positionnons nos travaux par rapport à l'état de l'art en présentant brièvement les résultats de la campagne MediaEval 2012. Nous comparons ensuite les trois systèmes de détection de violence que nous avons développés avec le système de l'équipe ARF [118, 61].

6.4.1 MediaEval Affect Task

6.4.1.1 Participation

L'année 2012 a montré que la tâche de détection de violence avait un intérêt pour le monde académique : 35 équipes de recherche ont manifesté un intérêt pour la tâche, et 11 d'entre elles se sont finalement inscrites, dont trois ayant déjà travaillé sur le sujet.

Au final, 8 équipes sont allées jusqu'au bout de la tâche et ont soumis leur résultats pour évaluation. Nous les présentons dans le tableau 6.6. L'équipe ARF est la seule a avoir proposé un système détectant des segments violents, ensuite agrégés sur des plans.

6.4.1.2 Résultats officiels

Les résultats officiels de l'édition 2012 sont montrés dans le tableau 6.7 et sur la figure 6.9. On distingue globalement trois groupes de participants, en termes de MAP@100. Un groupe de tête avec plus de 60 %, un peloton situé entre 30 et 50 % et les autres. En revanche, si l'on observe maintenant les courbes rappel-précision, la distinction est beaucoup moins claire. Il semble toutefois clair que les équipes ARF et SH sont loin devant tout le monde, et se talonnent l'une et l'autre pour des forts

Equipe	**Acronyme**	**Pays**
ARF	ARF	Autriche
DYNI - LSIS	DYNI	France
NII - Video Processing Lab	NII	Japon
Shanghai-Hongkong	SH	Chine
TUB - DAI	TUB	Allemagne
TUM	TUM	Allemagne-Autriche
LIG - MRIM	LIG	France
Technicolor*	TEC	France-UK

TABLEAU 6.6 – Équipes ayant franchi la ligne finale en 2012. L'équipe mentionnée par une étoile est une équipe formée des organisateurs de la tâche (notre équipe).

Equipe	**P**	**R**	**F1**	**MAP@100**	**MC**
ARF	31,24	66,15	42,44	**65,05**	3,56
DYNI	13,97	21,96	17,07	12,44	7,96
NII	11,17	96,50	20,02	30,82	1,28
SH	40,75	45,59	43,04	62,38	5,52
TUB	14,37	62,52	23,37	18,53	4,2
TUM	39,70	22,10	28,39	48,43	7,83
LIG	21,21	61,12	31,50	31,37	4,16
TEC*	21,72	67,27	32,83	61,82	3,56

TABLEAU 6.7 – Résultats officiels de chaque équipe pour leur meilleure soumission. P : précision, R : rappel, F1 : mesure F1

taux de rappel. Notre meilleur système est le système présenté section 6.3, et il se positionne troisième à la fois en terme de MAP@100 et en terme de précision-rappel. Le système basé sur la représentation TF-IDF que nous avons présenté en section 6.1 dans ce chapitre se classe $11^{\text{ème}}$ parmi les 35 expériences soumises par les participants.

Globalement, la campagne MediaEval 2012 nous a montré l'importance de l'intégration multimodale et temporelle, et nous a montré que des approches hiérarchiques naïves telles que celle présentée par l'équipe ARF peuvent donner des bons résultats. Comparativement à l'édition 2011, on a pu observer une réelle amélioration de la qualité des résultats obtenus par les différents participants.

6.4.2 Comparaison avec ARF

La figure 6.10 présente une comparaison de l'ensemble de nos systèmes développés dans cette partie avec celui de l'équipe ARF. Si les courbes précision-rappel obtenues par nos systèmes sont nettement inférieures à celles obtenues par ARF, il est intéressant de noter que notre meilleur système pendant la campagne, BN-SLP, est aussi le moins bon pour des taux de rappel élevés parmi nos trois systèmes. On note aussi que, toujours pour des taux de rappel élevés, notre système basé sur la représentation vectorielle par

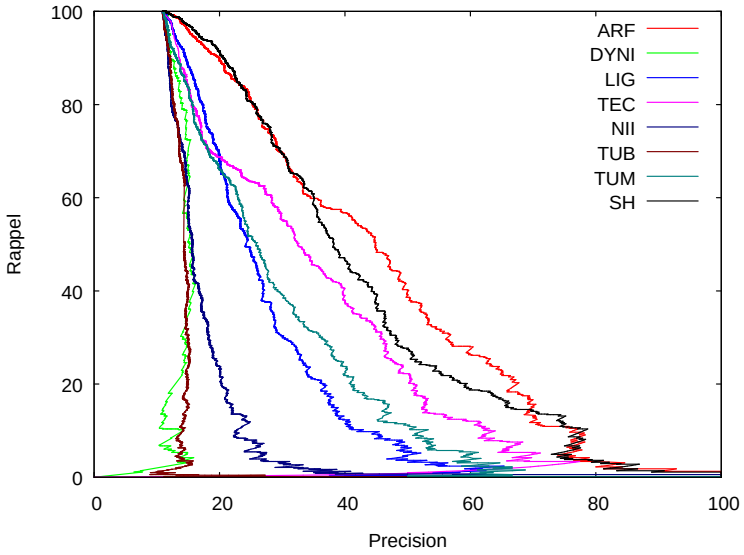


FIGURE 6.9 – Courbes rappel-précision des meilleures soumissions de chaque équipe.

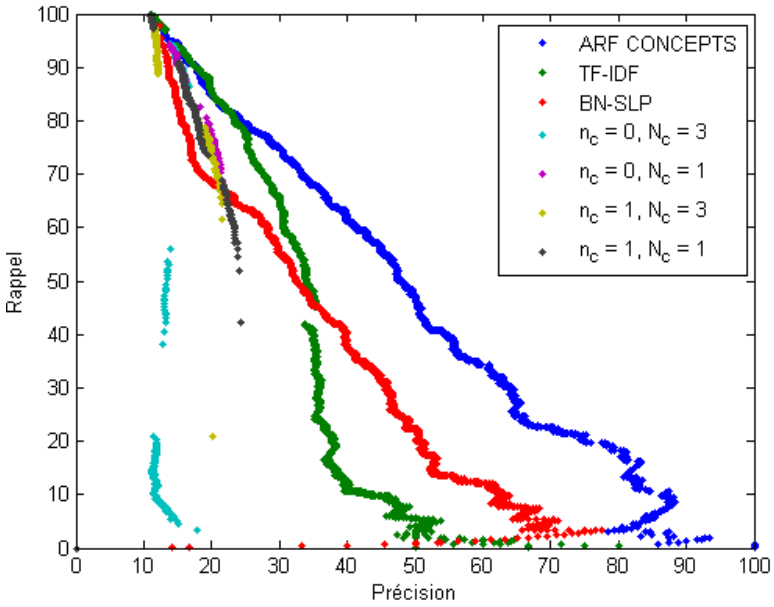


FIGURE 6.10 – Comparaisons de nos systèmes avec ARF. La courbe BN-SLP correspond au système multimodal.

sacs de mots TF-IDF est équivalent à celui proposé par ARF, bien qu'il soit basé uniquement sur l'audio.

En conclusion, les systèmes que nous avons présentés à MediaEval semblent être de bonnes pistes pour la détection de violence, et sont comparables au meilleur système de

l'état de l'art, c'est-à-dire celui de l'équipe ARF, pour des taux de rappel élevés. Etant donné qu'un rappel élevé indique que l'on détecte beaucoup de scènes violentes (sans donner d'indication sur les fausses alarmes), c'est aussi la zone qui correspond au cas d'usage originel. Ainsi, les résultats comparables à l'état de l'art dans cette zone sont une bonne indication quant à la qualité des systèmes que nous proposons. Nous pensons en revanche que les moins bons résultats des expériences utilisant un réseau bayésien hiérarchique sont dus à la trop faible quantité de concepts à l'étage conceptuel.

6.4.3 Vers MediaEval 2013

Si les résultats obtenus lors de l'édition 2012 de la campagne sont intéressants, la tâche en elle même souffre de plusieurs défauts :

- Les algorithmes sont comparés sur une base de test contenant trois films seulement, avec une métrique qui effectue une moyenne des résultats obtenus sur chacun des films. On peut donc se poser la question de la qualité de l'évaluation.
- La définition choisie n'est pas vraiment représentative du cas d'usage originel, car trop centrée sur le côté action. Certaines scènes très choquantes pour de jeunes enfants ne sont donc même pas annotées, car elles ne rentrent pas dans la définition.

Pour pallier ces défauts, l'édition 2013 introduit une nouvelle définition, cette fois plus subjective :

Définition 6.1. Nouvelle définition de la violence
Les évènements violents sont les scènes comportant de la violence physique ou des résultats de violence physique pouvant choquer un enfant de 8 ans. La violence psychologique et la violence verbale ne sont pas inclues.

Cette nouvelle définition aura l'avantage d'inclure des scènes contenant uniquement des cadavres, ou un homme en sang, comme dans *"Reservoir Dogs"*, sans montrer l'action violente ayant provoqué ces situations. Cela permettra aussi de ne pas forcément annoter certaines scènes, telles que celles où l'on peut observer un personnage qui boite, ou quelqu'un qui se coupe en se rasant.

La taille de l'ensemble de test passe aussi de trois à sept films, pour ainsi obtenir des résultats plus fiables. Une nouvelle sous-tâche facultative autorisant les participants à aller chercher des métadonnées externes aux DVD sera aussi mise en place.

Chapitre 7

Conclusions et perspectives

Nous avons présenté dans cette partie du mémoire nos travaux concernant la détection d'évènements sémantiques, avec application à la détection de violence. Nous rappelons dans ce chapitre nos principales contributions, avant de présenter des perspectives techniques à nos travaux.

7.1 Contributions

Notre première contribution a été de développer un jeu de données dédié à la détection de violence dans les films. Il est composé de 18 films Hollywoodiens, dans lesquels nous avons annoté la violence selon une définition que nous avons souhaitée la plus objective possible. Ce jeu de données a été développé dans le cadre d'une tâche de détection de violence que nous avons mise en place dans une campagne d'évaluation, appelée MediaEval Affect Task. Cette campagne nous a permis de proposer un cadre d'échange d'expériences scientifiques sur le sujet de la détection de violence pour permettre à la communauté de rechercher ce sujet, d'avancer de concert et de collaborer. Elle nous a aussi permis de comparer les systèmes que nous avons développés à l'état de l'art. Nous avons rendu les données publiques à l'adresse `https://research.technicolor.com/rennes/vsd/`.

Nous avons ensuite proposé trois systèmes de détection de violence. Nos deux premiers systèmes sont des systèmes monomodaux, utilisant uniquement le canal audio. Le premier système repose sur les représentations vectorielles par sac de mots audio TF-IDF, et nous permet de vérifier l'efficacité de ce type de représentations pour modéliser les plans vidéos des films. De plus, cela nous permet d'utiliser chaque mot audio extrait comme autant de concepts différents, mais cela a le désavantage de ne pas prendre en compte l'information temporelle contenue dans les séquences mots que nous avons développées dans la partie II. Ce système a obtenu la $11^{\text{ème}}$ place lors de la campagne MediaEval 2012, et une comparaison au système ARF montre que ces deux systèmes sont équivalents pour des taux de rappel élevés, correspondant au cas d'usage de la campagne MediaEval.

Pour prendre en compte l'information temporelle, inspirés par l'équipe ARF, nous

avons mis en place un système basé sur l'utilisation de réseaux bayésiens contextuels hiérarchiques naïfs, nous permettant d'utiliser les détecteurs de concepts développés dans le chapitre 3. Ce nouveau système nous permet d'utiliser directement le détecteur de concepts comme entrée d'un détecteur de violence. Nous pensons que les moins bons résultats que nous obtenons avec ce système sont principalement à imputer au fait que nous utilisons seulement deux concepts dans notre détecteur, coups de feu et explosions, qui ne représentent pas la majorité des scènes violentes annotées.

Enfin, nous présentons un système multimodal, grâce auquel nous étudions l'apprentissage de structure dans les réseaux bayésiens, l'intégration temporelle et l'intégration multimodale audio-vidéo à l'aide d'attributs audio et vidéos bas niveau. Nous montrons notamment que l'apprentissage de structure K2, par exemple, nous permet d'obtenir des graphes facilement interprétables, nous permettant aussi de séparer les attributs liés à la violence, de ceux qui ne le sont pas. Nous montrons aussi l'importance de l'intégration multimodale et l'importance de l'intégration temporelle, ces combinaisons étant celles donnant les meilleurs résultats. En termes de métrique officielle de la campagne, ce système est celui qui a obtenu les meilleurs résultats parmi ceux que nous présentons, décrochant la troisième place en 2012, avec, au final, un taux de MAP@100 de $\simeq 62\,\%$.

7.2 Perspectives

Les systèmes que nous avons développés mettent en lumière de nombreuses perspectives à nos travaux. Nous présentons celles qui nous semblent les plus importantes dans cette section, système par système.

Utilisation des séquences de mots audio : Nos deux premiers systèmes monomodaux, le système TF-IDF et le système à réseau bayésien hiérarchique contextuel, utilisent tous les deux les séquences de mots audio que nous avons définies dans la partie II sur la détection de concepts audio dans les films. Or, dans cette partie sur la détection de violence, nous avons seulement utilisé les MFCC avec un quantifieur dans le système TF-IDF, trois quantifieurs pour le réseau bayésien hiérarchique contextuel. Nous pensons qu'utiliser les autres artifices que nous avons définis précédemment, tels que l'utilisation d'autres types d'attributs ou de l'analyse factorielle, permettrait d'améliorer les résultats.

TF-IDF : La pondération des mots utilisée dans la section 6.1 a été développée dans le cadre de la recherche de documents textuels similaires, et ne prend pas en compte la longueur des segments considérés ainsi que la longueur des plans vidéos. Prendre en compte ces durées dans la pondération permettrait de renforcer l'importance des segments et des plans courts. On peut aussi imaginer que l'on puisse développer une pondération prenant directement en compte l'information temporelle contenue dans les plans vidéos. Enfin, l'utilisation d'autres classifieurs tels que les réseaux bayésiens peut aussi être envisagée.

Réseaux bayésiens hiérarchiques : Le système développé section 6.2 utilise seulement deux concepts, les coups de feu et les explosions. L'utilisation de concepts supplémentaires, tels que les cris, permettrait certainement d'améliorer la représentation du flux audio. Cela permettrait ainsi probablement d'améliorer les résultats de détection de violence. On pourrait même envisager d'utiliser des nœuds à valeurs continues de façon à pouvoir utiliser les sorties des détecteurs de concepts pour l'apprentissage du détecteur de violence, permettant ainsi d'adapter complètement l'idée de l'équipe ARF, et aussi peut-être de limiter les problèmes d'utilisation de mémoire vive liés au type de réseau bayésien que nous proposons d'utiliser.

Apprentissage de structure : L'apprentissage de structure que nous avons réalisé dans la section 6.3 est un apprentissage descriptif, c'est-à-dire qu'il optimise la probabilité jointe des variables d'entrée et de sortie du système $P(C, \mathbf{X})$, sans faire de distinction particulière entre les variables et le nœud classe. Les résultats sont bons, mais cela ne correspond pas à un système de classification, dont le but est d'obtenir la classe de l'échantillon sachant les valeurs observées des variables d'entrées, c'est-à-dire $P(C|\mathbf{X})$. Pour cela l'utilisation d'un critère discriminant pourrait être intéressant, à l'image de celui développé par [58, 57]. En revanche, les graphes obtenus seraient plus difficilement interprétables, ne permettant pas de tirer des conclusions sur les dépendances entre les attributs.

Multimodalité : Dans le système multimodal, nous nous sommes concentrés sur les modalités audio et vidéo. Il serait intéressant d'étendre le système à la modalité textuelle, soit en incluant les sous-titres, soit grâce à l'utilisation du script du film si celui-ci est disponible. L'utilisation de métadonnées externes est aussi envisageable. Ces informations pourraient par exemple être utilisées comme *a priori* sur les scènes violentes.

MediaEval Affect Task : Comme indiqué dans la section 6.4.3, la tâche que nous avons développée dans le cadre de MediaEval sur la détection de violence dans les films est amenée à évoluer. Tout d'abord, la taille de l'ensemble de test va augmenter dans l'édition 2013, de sorte à obtenir des résultats plus significatifs. Une définition supplémentaire va aussi être proposée, plus subjective et plus adaptée au cas d'utilisation à l'origine de la tâche. Les participants auront aussi la possibilité de soumettre une "bande annonce" des scènes les plus violentes.

Dans un futur plus éloigné, nous avons aussi pensé à ouvrir cette campagne à d'autres types de contenus, tels que les flux télévisuels ou les vidéos YouTube. Dans le but de pouvoir comparer plus efficacement les systèmes, nous pensons aussi imposer certaines configurations des systèmes. Nous pouvons ainsi imaginer fixer une modalité ou une combinaison de modalités, ou encore fournir les attributs aux participants. Cela nous permettrait ainsi de vraiment comparer les systèmes en se focalisant soit sur les attributs, soit sur les modalités, soit sur les techniques, ...

Conclusions et perspectives générales

Dans cette thèse, nous nous sommes intéressés à deux types de problèmes : la détection d'évènements audio et la détection d'évènements sémantiques violents dans les films. Le détail des contributions que nous avons apportées ainsi que les perspectives que nous envisageons pour chacun de ces problèmes sont présentés dans les chapitres 4 et 7 respectivement. Dans ce chapitre, nous rappelons brièvement nos principales contributions avant de présenter des perspectives générales de recherche à nos travaux.

Contributions principales

Notre première contribution est probablement d'avoir mis en place un jeu de données dédié à la détection de violence que nous avons rendu public. Ce jeu de données est composé de 18 films. Sur ces 18 films, 15 sont réservés à l'apprentissage des systèmes, et 3 sont utilisés pour les tests. Les 15 films d'apprentissage sont annotés en terme de violence, ainsi qu'en terme de sept concepts vidéo et trois concepts audio. Les trois films de test sont uniquement annotés en terme de violence. Ce jeu de données a été développé dans le cadre de la mise en place d'une tâche dédiée à la détection de violence dans la campagne d'évaluation MediaEval. Nous avons contribué à la définition et l'organisation de cette tâche entre 2011 et 2013.

Nous avons montré dans la première partie de ce mémoire que le fait de travailler avec des films induisait une divergence statistique entres les bandes sonores des films. Nous pensons que cette divergence statistique est due aux post-traitements audio effectués lors de la création du film, ainsi qu'à la complexité des bandes sonores. Cette divergence statistique induit un problème de généralisation, qui a pour effet de pénaliser très fortement la qualité des résultats obtenus.

Nous avons ensuite apporté des éléments pour la résolution de ce problème de généralisation. Nous avons proposé d'utiliser des séquences de mots audio, au travers de la quantification par partie, permettant ainsi d'obtenir plusieurs mots par échantillon audio. Nous avons aussi proposé d'utiliser des réseaux bayésiens contextuels pour classifier les séquences de mots audio ainsi obtenues. Nous montrons que cette technique nous permet d'obtenir des résultats comparables à l'état de l'art sur les mêmes données. Nous montrons aussi que les différents types d'attributs que nous avons utilisés

semblent complémentaires par rapport aux évènements que nous avons tentés de détecter, à savoir les coups de feu et les explosions, et nous montrons que notre système peut être soit très robuste au seuil de décision, soit proposer un grand nombre de points de fonctionnement en fonction de la configuration dans laquelle il est utilisé. Nous améliorons ensuite fortement les résultats obtenus avec les attributs MFCC, notamment leur capacité à détecter les explosions, en adaptant une technique d'analyse factorielle développée dans le cadre de la reconnaissance de locuteur. Cette technique nous permet de modéliser la variabilité entre les films et de la compenser directement dans les attributs. Nous montrons qu'en utilisant l'analyse factorielle sur les MFCC avant de calculer le dictionnaire de mots audio, il est possible d'améliorer les résultats de détection de concepts.

Dans la deuxième partie, nous nous sommes intéressés à la détection d'évènements sémantiques dans les films, avec application à la détection de violence. Nous proposons ainsi trois pistes de recherche pour ce problème : deux pistes audio, et une piste multimodale. Nous proposons dans un premier temps d'utiliser une représentation de type TF-IDF des plans vidéos, en l'appliquant aux séquences de mots audio extraites dans la première partie. Ce système revient à considérer les mots audio comme des concepts non-supervisés. Les résultats que nous obtenons sont bons, et nous montrons que ce système est équivalent voire meilleur que le système de l'équipe ARF, vainqueur de MediaEval 2012, pour des taux de rappel élevés. Pour essayer de prendre en compte l'évolution temporelle des mots audio dans notre système, nous avons ensuite adapté le système développé dans la première partie en utilisant les détecteurs de concepts audio comme entrées du détecteur de violence, et nous avons mis en place ce système en utilisant des réseaux bayésiens contextuels hiérarchiques. Les résultats que nous obtenons sont bien moins bons que pour le système précédent mais restent encourageants étant donné le faible nombre de concepts inclus dans notre détecteur de concepts. Enfin, nous avons montré l'intérêt des représentations multimodales, et de l'intégration temporelle grâce à notre dernier système. Ce système nous a permis entre autres de montrer le grand intérêt de l'apprentissage de structure pour expliquer les liens entre les différents attributs. Ce dernier système a obtenu la troisième place lors de la campagne MediaEval 2012.

Perspectives générales

Les travaux que nous avons présentés nous amènent de nombreuses perspectives, et nous choisissons ici de présenter celles que nous avons jugées les plus intéressantes.

Tout d'abord, il serait intéressant d'appliquer ces recherches à d'autres types d'évènements, tels que les cris ou les émotions, ou d'autres types de contenus. On pourrait par exemple appliquer ces recherches à la caractérisation de la violence dans les bandes annonces. Il arrive souvent, en allant voir un film au cinéma avec des enfants, de devoir visualiser une bande annonce totalement inadaptée au niveau de violence du public du film. Nous pourrions ainsi développer une mesure de violence globale sur les bandes annonces et sur les films, en utilisant les systèmes que nous avons développés. Cela nous

permettrait ainsi de proposer aux cinémas un système les aidant à créer leurs séquences de bandes annonces en fonction du public du film.

Dans une autre direction, la campagne MediaEval Affect 2012 ainsi que notre système nous ont appris à quel point il est important de considérer toutes les sources d'information d'un film pour détecter des concepts sémantiques, en particulier la violence. Cela nous montre à quel point chercher des solutions multimodales à un problème multimodal peut être bénéfique. De plus, nous pensons que l'on peut construire un bon détecteur de violence en se basant sur des détecteurs de concepts de faible performance. Nous pourrions donc appliquer ce principe à un système multimodal, en utilisant d'autres types d'évènements. On pourrait aussi imaginer détecter des évènements audio visuels, par la construction d'un dictionnaire audio visuel tel que celui décrit par Irie *et al.* [62] par exemple.

Enfin, en observant les expériences audio menées, nous remarquons que l'expérience basée sur les réseaux bayésiens contextuels hiérarchiques naïfs (CHBN) donne de moins bons résultats que le système fondé sur la représentation TF-IDF, mais qu'il donne une meilleure segmentation du flux audio. De plus, on peut aussi remarquer que la représentation TF-IDF revient à utiliser les mots audio directement comme des concepts. Enfin, il est essentiel de comprendre que l'annotation des films en termes de concepts est une activité chronophage, et très dépendante du film que l'on annote : par exemple annoter les coups de feu dans *"Saving Private Ryan"* ne nécessite pas du tout le même effort que dans *"Billy Elliot"*.

Ces remarques nous amènent à penser que la meilleure direction de recherche pour la détection de concepts, ou même dans le but de décrire la bande son d'un film, passe probablement par l'acquisition non supervisée de concepts pour décrire le signal audio. Ainsi, cela nous permettrait de ne pas nous limiter aux seuls concepts annotés mais de les découvrir nous-mêmes. Nous pensons que cela aurait vraiment un grand effet sur la résolution du problème de généralisation. Pour arriver à faire cela, nous avons identifié deux pistes de recherche, qui ne sont pas indépendantes l'une de l'autre.

La première piste que nous avons identifiée consiste à utiliser des variables latentes dans le réseau bayésien. Si l'on prend l'exemple du CHBN, on pourrait remplacer l'étage conceptuel par un étage de variables latentes. Les variables latentes dans les réseaux bayésiens ne sont pas un sujet nouveau. On peut citer les travaux de Zhang *et al.* [76, 137, 138] ou de Langseth *et al.* [81, 82] sur la découverte de variables latentes dans les réseaux bayésiens. Globalement, l'idée est d'insérer une variable dont on ne connaît rien quelque part dans le réseau et de vérifier que son insertion permet d'améliorer l'adaptation du réseau aux données. Le problème de l'utilisation des variables latentes est de connaître leur nombre, leur taille et leur position optimale dans le réseau permettant d'améliorer sa vraisemblance. Les principaux algorithmes existants résolvent ce problème par des méthodes gloutonnes souvent coûteuses lorsque l'on considère que les paramètres de ces variables doivent être estimés à l'aide d'un algorithme EM par exemple. Il y a donc beaucoup de "place" disponible ici pour tenter d'optimiser ce problème.

La deuxième piste à laquelle nous avons pensé est de continuer dans le sens que nous avons pris dans ce mémoire, en ajoutant une segmentation supplémentaire, permettant

de séparer les différents évènements sonores, ou unités de son, telle que la segmentation par front d'attaque proposée par Bello *et al.* [7]. Associé à la segmentation en segments stationnaires que nous avons utilisée dans ce mémoire, nous aurions ainsi des unités de son décomposées en segments stationnaires successifs. Nous pourrions alors réfléchir à une méthode temporelle permettant de grouper les unités de son similaires ensemble, telle que la méthode de groupement par HMM proposée par Kumar *et al.* [79]. Nous pourrions ainsi grouper les évènements audio similaires et construire des détecteurs de ces évènements que nous utiliserions ensuite dans le contexte des CHBN. Cela aurait l'avantage de nous offrir plus de concepts nous permettant de décrire la bande sonore des films, et probablement d'améliorer la détection de violence.

L'acquisition non-supervisée de concepts pourrait aussi s'appliquer aux autres modalités des contenus vidéos. On obtiendrait ainsi plusieurs séquences d'évènements permettant de décrire le contenu. Ces séquences pourrait ensuite être utilisées pour décrire des évènements sémantiques, telles que la violence ou les émotions, en tentant d'arranger, et de synchroniser, les résultats obtenus.

Annexe A

Preuves mathématiques liées à l'analyse factorielle

Nous proposons dans ce chapitre de montrer comment les équations des facteurs locuteur et canal liés à l'analyse factorielle sont dérivées. Nous proposons aussi de montrer l'existence de la solution des équations des facteurs, ainsi que de la matrice **U**.

A.1 Détermination des facteurs locuteur et canal

Nous souhaitons montrer dans un premier temps comment obtenir l'équation donnant la valeur des facteurs **q** dans l'équation 3.4.

Preuve : Comme indiqué dans la section 3.3.1.3, la solution par *maximum a posteriori* permettant de trouver **q** est équivalente à estimer les facteurs **q** de sorte qu'ils maximisent :

$$\max_{\mathbf{q}} p(\mathbf{X}|\mathbf{q},\boldsymbol{\mu},\mathbf{G})\mathcal{N}(\mathbf{q}|\mathbf{0},\mathbf{I})$$

En considérant une seule composante g pour le mélange de gaussiennes, et en remplaçant p et $\mathcal{N}$ par leurs expressions analytiques, on obtient :

$$\begin{aligned} p(\mathbf{X}|\mathbf{q},\boldsymbol{\mu},\mathbf{G})\mathcal{N}(\mathbf{q}|\mathbf{0},\mathbf{I}) = &\sum_{t=1}^{T} \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu})\right) \\ &* \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{1}{2}\mathbf{q}^T\mathbf{q}\right) \end{aligned} \tag{A.1}$$

Ainsi, le logarithme de cette expression devient :

$$\log\left(p(\mathbf{X}|\mathbf{q},\boldsymbol{\mu},\mathbf{G})\mathcal{N}(\mathbf{q}|\mathbf{0},\mathbf{I})\right) = CST - \sum_{t=1}^{T}\frac{1}{2}(t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu}) - \frac{1}{2}\mathbf{q}^T\mathbf{q} \tag{A.2}$$

où :

$$CST = \log\left(\sum_{t=1}^{T}\frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\frac{1}{(2\pi)^{\frac{D}{2}}}\right) \text{ est une constante.} \tag{A.3}$$

Le problème de maximisation revient donc à minimiser :

$$\begin{aligned}\min_{\mathbf{q}} \sum_{t=1}^{T}(t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu}) + \mathbf{q}^T\mathbf{q} = &\sum_{t=1}^{T}(t-\boldsymbol{\mu_{prior}})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu_{prior}}) \\ &- 2\sum_{t=1}^{T}(\mathbf{Gq})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu_{prior}}) \\ &+ \sum_{t=1}^{T}(\mathbf{Gq})^T\boldsymbol{\Sigma}^{-1}\mathbf{Gq} \\ &+ \mathbf{q}^T\mathbf{q} \end{aligned} \tag{A.4}$$

Ainsi, dans l'équation précédente, il reste à estimer $\sum_{t=1}^{T} 1$ et $\sum_{t=1}^{T} t$. Pour cela, il suffit de les remplacer par leurs espérances mathématiques, correspondant à l'occupation statistique n_g pour la première somme, et la somme pondérée des échantillons $\mathbf{S}_{\mathbf{X},g}$ pour le deuxième. Avec ces définitions, il est possible de réécrire :

$$\begin{aligned}\sum_{t=1}^{T}(t-\boldsymbol{\mu_{prior}}) &= \sum_{t=1}^{T} t - \sum_{t=1}^{T}\boldsymbol{\mu_{prior}} \\ &= S_{\mathbf{X},g} - n_g\boldsymbol{\mu_{prior}} \\ &= S_{\mathbf{X},g|\boldsymbol{\mu_{prior}}}\end{aligned}$$

On obtient par conséquent :

$$\begin{aligned}\sum_{t=1}^{T}(t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu}) + \mathbf{q}^T\mathbf{q} = &\sum_{t=1}^{T}(t-\boldsymbol{\mu_{prior}})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu_{prior}}) \\ &- 2(\mathbf{Gq})^T\boldsymbol{\Sigma}^{-1}S_{\mathbf{X},g|\boldsymbol{\mu_{prior}}} \\ &+ (\mathbf{Gq})^T n_g\boldsymbol{\Sigma}^{-1}\mathbf{Gq} \\ &+ \mathbf{q}^T\mathbf{q} \end{aligned} \tag{A.5}$$

L'équation A.5 peut se réécrire sous la forme d'une équation quadratique :

$$\sum_{t=1}^{T}(t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu}) + \mathbf{q}^T\mathbf{q} = \mathbf{q}^T\mathbf{Aq} + \mathbf{q}^T\mathbf{b} + \mathbf{c} \tag{A.6}$$

où :

$$A = \mathbf{I} + \mathbf{G}^T n_g\boldsymbol{\Sigma}^{-1}\mathbf{G} \tag{A.7}$$

$$b = -2\mathbf{G}^T\boldsymbol{\Sigma}^{-1}S_{\mathbf{X},g|\boldsymbol{\mu_{prior}}} \tag{A.8}$$

$$c = \sum_{t=1}^{T}(t-\boldsymbol{\mu_{prior}})^T\boldsymbol{\Sigma}^{-1}(t-\boldsymbol{\mu_{prior}}) \tag{A.9}$$

La solution pour $\mathbf{q}$ est donc triviale, dans le cas où $\mathbf{A}$ est inversible :

$$\begin{aligned}\mathbf{q} &= -\frac{\mathbf{b}}{2\mathbf{A}} \\ &= \mathbf{A}^{-1}\mathbf{G}^T\boldsymbol{\Sigma}^{-1}S_{\mathbf{X},g|\boldsymbol{\mu_{prior}}}\end{aligned}$$

ce qui correspond aux valeurs que l'on cherchait pour $\mathbf{z}$ et $\mathbf{q}$. L'extension au cas où il y a plusieurs gaussiennes, c'est-à-dire $g > 1$, est triviale, car il est possible de concaténer les matrices. On obtient ainsi les mêmes résultats en utilisant le supervecteur de moyennes et la supermatrice de variance, et en diagonalisant les statistiques d'ordre zéro et de premier ordre.

⊠

A.2 Inversibilité de la matrice A

Le résultat précédent suppose que la matrice $\mathbf{A}$ est inversible, ce qui nous permet de prouver que $\mathbf{q}$ a toujours une solution. Démontrons donc a présent que $\mathbf{A}$ est inversible.

Preuve : Pour démontrer que $\mathbf{A}$ est inversible, nous démontrons qu'elle est définie positive, c'est-à-dire :

$$\mathbf{x}^T\mathbf{A}\mathbf{x} > 0 \text{ pour tout } \mathbf{x} \neq 0$$

En insérant l'expression de $\mathbf{A}$, on obtient donc :

$$\begin{aligned}\mathbf{x}^T\mathbf{A}\mathbf{x} &= \mathbf{x}^T\left(\mathbf{I} + \mathbf{G}^T n_g \boldsymbol{\Sigma}^{-1}\mathbf{G}\right)\mathbf{x} \\ &= ||\mathbf{x}||^2 + \mathbf{x}^T\left(\mathbf{G}^T n_g \boldsymbol{\Sigma}^{-1}\mathbf{G}\right)\mathbf{x} \end{aligned} \tag{A.10}$$

En définissant l'ensemble de vecteurs $\mathcal{E} := \{e_i = \sqrt{n_g\boldsymbol{\Sigma}^{-1}}g_i, i \in [0, R_G]\}$, où g_i correspond aux lignes de la matrice G, on montre que la matrice $\mathbf{A}' = \mathbf{G}^T n_g \boldsymbol{\Sigma}^{-1}\mathbf{G}$ représente la matrice de Gram de l'ensemble $\mathcal{E}$, c'est-à-dire que ses entrées correspondent au produit scalaire de ses éléments $\mathbf{A}'_{ij} =< e_i, e_j >$. Or, les matrices de Gram sont définies semi-positives par construction, c'est-à-dire :

$$\mathbf{x}^T\mathbf{A}'\mathbf{x} \geq 0 \text{ pour tout } \mathbf{x} \neq 0$$

On a donc :

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \underbrace{||\mathbf{x}||^2}_{>0} + \underbrace{\mathbf{x}^T\left(\mathbf{G}^T n_g \boldsymbol{\Sigma}^{-1}\mathbf{G}\right)\mathbf{x}}_{\geq 0} > 0 \tag{A.11}$$

Ainsi, on obtient bien $\mathbf{A}$ définie positive, ce qui indique que la matrice est inversible.

⊠

A.3 Existence de la solution de la matrice U

On a vu dans la section 3.3.1.4 que la matrice **U** est définie par :

$$\mathbf{U}_g \mathbf{A}_g = \mathbf{B}_g \tag{A.12}$$

où

$$\begin{aligned} \mathbf{A}_g &= \sum_{s=1}^{S} \sum_{h=1}^{H(s)} n_g(h,s) \left(\mathbf{z}(h,s)\mathbf{z}(h,s)^T + \mathbf{A}(h,s)^{-1} \right) \\ \mathbf{B}_g &= \sum_{s=1}^{S} \sum_{h=1}^{H(s)} \mathbf{S}_{\mathbf{X}(s,h),g|\boldsymbol{\mu}(s)} \mathbf{z}(h,s)^T \end{aligned}$$

Il nous faut donc montrer l'inversibilité de $\mathbf{A}_g$ pour démonter l'existence de la solution de **U**.

Preuve : Encore une fois, il nous suffit de démonter que A_g est définie positive. La solution est triviale : on a $\mathbf{z}(h,s)\mathbf{z}(h,s)^T$ définie positive par construction et **A** définie positive. La matrice A_g représente donc la somme de matrices définies positives par construction, on obtient par conséquent que A_g est définie positive, et donc inversible.

⊠

Annexe B

Besoin en mémoire du réseau contextuel hiérarchique naïf

Nous proposons dans cette annexe d'expliquer pourquoi le réseau bayésien contextuel hiérarchique (CHBN) naïf peut poser un problème de mémoire vive lors de son apprentissage et de son utilisation. Nous proposons dans cette annexe d'estimer les besoins en mémoire du CHBN et montrer l'influence des différents paramètres.

B.1 Besoin en mémoire d'un réseau bayésien

D'une manière générale, l'espace mémoire requis par réseau bayésien est fortement corrélé, voire proportionnel au nombre de paramètres qui le caractérisent. Pour rappel, un BN modélise la probabilité jointe d'un ensemble de variables $\mathbf{X}$ en simplifiant le calcul à l'aide des dépendances connues ou estimées entre les différentes variables. Le calcul de la probabilité jointe est donc défini par l'équation 1.4, et chaque nœud du graphe G contient la quantité $p_G(x_i|Pa_G(x_i))$. Dans le cas où les variables sont discrètes, les paramètres peuvent être représentés par une table de vérité et le nombre de paramètres correspond au nombre d'états du nœud considéré multiplié par le nombre d'états des parents du nœud. Dans le cas où les variables sont continues, les paramètres correspondent aux paramètres de la distribution représentant les données.

Nous nous intéressons dans cette annexe au cas où les variables sont discrètes. Il nous faut donc estimer le nombre de paramètres pour avoir une estimation du besoin mémoire. C'est ce que nous faisons dans la section suivante.

B.2 Estimation du nombre de paramètres d'un réseau bayésien hiérarchique contextuel naïf

La figure B.1 présente un exemple de CHBN, en précisant à chaque étage les variables nécessaires à la définition du nombre de paramètres. Nous définissons dans cette section le nombre de paramètres nécessaires à chaque étage du réseau CHBN.

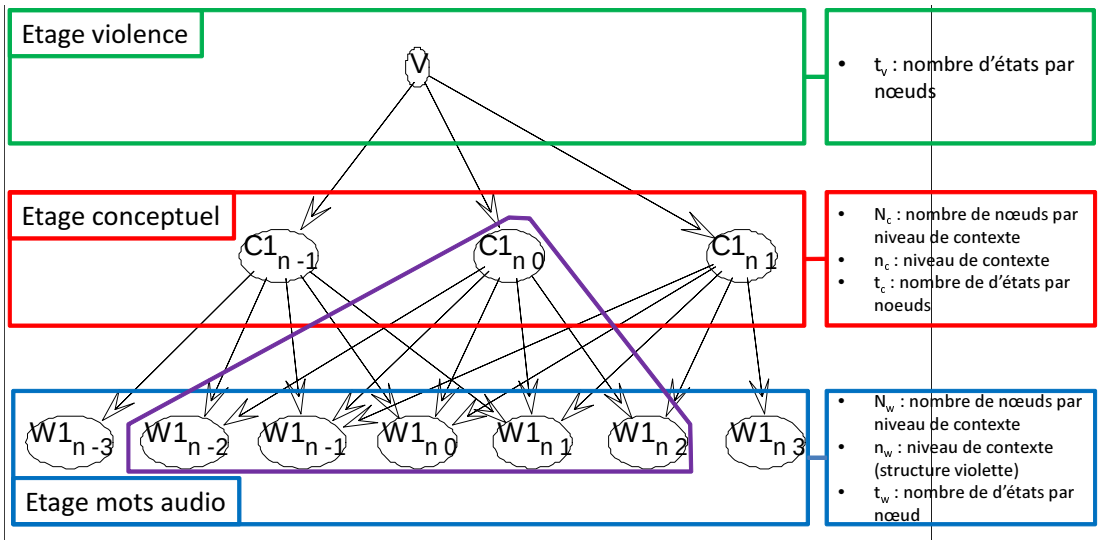


FIGURE B.1 – Exemple de réseau bayésien contextuel hiérarchique naïf et les variables associées à chaque étage.

B.2.1 Étage violence

Cet étage étant composé d'un seul nœud, le nombre de paramètres correspond au nombre d'états de la variable, t_v.

B.2.2 Étage conceptuel

A cet étage, chaque nœud est connecté au nœud de l'étage violence. Ainsi, si chaque concept a t_c états, le nombre de paramètres de chaque nœud est donc : $t_c * t_v$. Le nombre de nœuds correspond au produit du nombre de niveaux de contexte $2n_c+1$ par le nombre de nœuds concepts utilisés par niveau de contexte N_c. Le nombre total de paramètres à cet étage est donc :

$$M_c(t_v, t_c, N_c, n_c) = t_c * t_v * N_c * (2n_c + 1)$$

B.2.3 Étage mots audio

Cet étage est plus compliqué à appréhender car le nombre de connexions de chaque nœud varie en fonction du niveau temporel, et du nombre de concepts N_c par niveau de contexte. En définissant N_w le nombre de mots par niveau de contexte, t_w le nombre d'états associé à chaque nœud et n_w le niveau de contexte de cet étage, correspondant au niveau de contexte de la structure entourée en violet sur la figure B.1, on définit le nombre maximal de niveaux temporels de concepts auxquels un mot peut être connecté par $TCM = (2\min(n_w, n_c) + 1)$ et le nombre de connexions maximum d'un nœud par $CM = N_c * TCM$. On définit aussi le nombre de mots ayant le maximum de connexions par $W_{CM} = N_w * (2|n_w - n_c| + 1)$. En observant le réseau, on remarque que le nombre

maximum de connections est toujours observé pour les nœuds au centre de la structure, et qu'ensuite le nombre de connexions temporelles décrémente de 1 à chaque niveau temporel jusqu'à atteindre une seule connection temporelle aux extrêmes du réseau. Dans l'exemple de la figure B.1, cela veut dire que $W1_{n-1}$, $W1_{n0}$ et $W1_{n1}$ ont trois connexions chacun, $W1_{n-2}$ et $W1_{n2}$ deux connexions et $W1_{n-3}$ et $W1_{n3}$ une seule.

Le nombre de paramètres de chaque nœud est aussi défini par $p_w = t_w * t_c^{N_c * C}$, où C est le nombre de connexions du nœud. Le nombre total de paramètres à cet étage est donc :

$$\begin{aligned} M_w(t_c, N_c, n_c, t_w, n_w, n_c) &= N_w * (2|n_w - n_c| + 1) * t_w * t_c^{N_c*(2\min(n_c,n_w)+1)} \\ &\quad + 2 \sum_{i=1}^{2\min(n_c,n_w)} N_w * t_w * t_c^{N_c*i} \\ &= W_{CM} * t_w * t_c^{CM} \\ &\quad + 2 \sum_{i=1}^{TCM-1} N_w * t_w * t_c^{N_c*i} \end{aligned}$$

B.2.4 Nombre total de paramètres

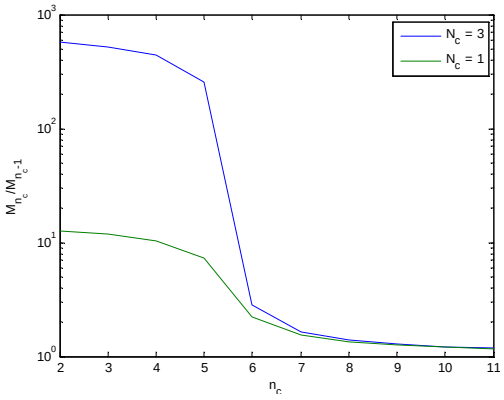
Le nombre total de paramètres du CHBN est donc défini au final par :

$$\boxed{M(t_v, N_c, N_w, t_c, t_w, n_w, n_c) = t_v + M_c(t_v, t_c, N_c, n_c) + M_w(t_c, N_c, n_c, t_w, n_w, n_c)} \tag{B.1}$$

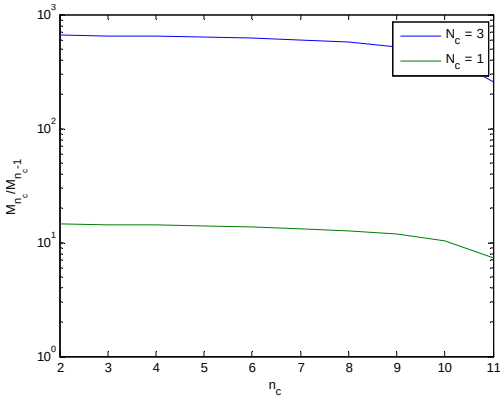
On peut donc maintenant s'intéresser aux besoins en mémoire du réseau en fonction des paramètres, et surtout à la variation de ce besoin en fonction des différentes configurations. Nous nous intéressons donc à la variation du nombre de paramètres en fonction de n_c, n_w, N_c et N_w. Les paramètres par défaut utilisés sont : $t_v = 2$, $t_c = 3$ si $N_c > 1$, $t_c = 4$ si $N_c = 1$, $n_c = 1$, $t_w = 130$, $N_w = 3$ et $n_w = 5$. Ces valeurs par défaut correspondent à peu près aux valeurs utilisées dans la section 6.2 du chapitre 6. Les résultats sont présentés sur les figures B.2a, B.2b, B.2c, B.2d et B.2e.

Ainsi, il apparaît clairement, en observant notamment les figures B.2a, B.2b, B.2c et B.2d, que les variables n_c et n_w ont exactement la même influence. Cela se comprend en observant la formule, car elles régissent le nombre maximum de connexions des nœuds de l'étage mots audio ainsi que le nombre de mots ayant un nombre de connexions maximal. L'importante diminution de la variation observée sur la figure B.2a autour de $n_c = 6$ s'explique par le fait que le nombre maximal de connexions des nœuds ne varie plus. Dans le cas de la figure B.2c, on a directement $n_w > n_c$, ce qui explique les valeurs obtenues. Enfin, il est aisé de comprendre que N_c a une bien plus grande influence que N_w, car N_c affecte directement le nombre de connexions, contrairement à N_w.

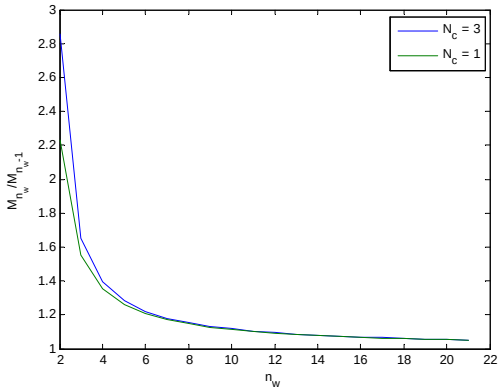
Comprendre pourquoi, lors des expériences de la section 6.2 du chapitre 6, on ne peut pas obtenir de résultats pour $N_c = 3$ et $n_c > 1$ devient donc intuitif : les besoins en mémoire sont multipliés par environ 600. Si l'expérience prend 4 Go de mémoire vive avec $n_c = 1$, alors, pour $n_c = 2$, il faudrait 2,4 To de mémoire vive environ.



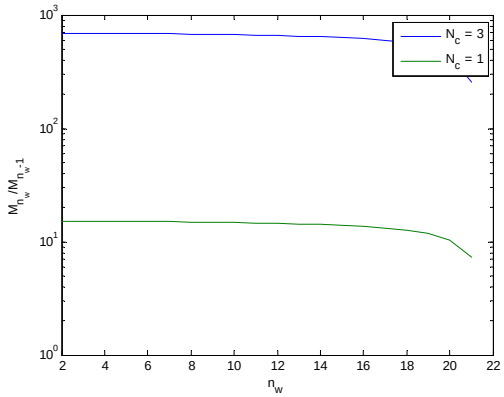
(a) Variation en fonction de n_c, $n_w = 5$.



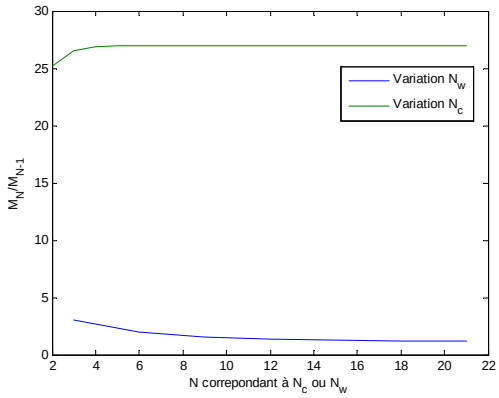
(b) Variation en fonction de n_c, $n_w = 11$.



(c) Variation en fonction de n_w, $n_c = 1$.



(d) Variation en fonction de n_w, $n_c = 21$.



(e) Variation en fonction de N_c et N_w.

FIGURE B.2 – Variation du besoin en mémoire en fonction des paramètres.

Ce problème pourrait par exemple être résolu en diminuant le nombre de paramètres du réseau (en utilisant des distributions continues sur certains nœuds), ou en le coupant

en deux : un réseau pour détecter les concepts, avec un seul niveau de contexte sur les concepts, puis un deuxième réseau utilisant les concepts, pour détecter la violence. Les sorties du détecteur de concepts pourraient ainsi être utilisées comme attributs du détecteur de violence.

Bibliographie

[1] Violence : a Public Health Priority. Technical report, World Health Organization, Geneva, Switzerland, 1996.

[2] R. André-Obrecht. A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36 :29–40, 1988.

[3] P.K. Atrey, M.A. Hossain, A. El Saddik, and M.S. Kankanhalli. Multimodal Fusion for Multimedia Analysis : a Survey. *Multimedia Systems*, 16 :345–379, 2010.

[4] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio Based Event Detection for Multimedia Surveillance. In *Proceedings of the 31st IEEE International Conference on Acoustics, Speech and Signal Processing*, 14-19 2006.

[5] S. Baghdadi. *Extraction Multimodale de Métadonnées de Séquences Videos dans un Cadre Bayésien*. PhD thesis, Université de Rennes 1, 2010.

[6] Y. Baveye, F. Urban, C. Chamaret, V. Demoulin, and P. Hellier. Saliency-Guided Consistent Color Harmonization. In Shoji Tominaga, Raimondo Schettini, and Alain Trémeau, editors, *Proceedings of the 4th Computational Color Imaging Workshop*, volume 7786 of *Lecture Notes in Computer Science*, pages 105–118. Springer Berlin Heidelberg, 2013.

[7] J.P. Bello, C. Duxbury, M. Davies, and M. Sandler. On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain. *Signal Processing Letters*, 11(6) :553–556, 2004.

[8] K.P. Bennett and C. Campbell. Support Vector Machines : Hype or Hallelujah ? *SIGKDD Explorations Newsletter*, 2(2) :1–13, 2000.

[9] J. Bilmes. Dynamic Bayesian Multinets. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 38–45, 2000.

[10] C.M. Bishop and M.E. Tipping. A Hierarchical Latent Variable Model for Data Visualization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3) :281 –293, 1998.

[11] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason. ALIZE/SpkDet : A State-of-the-Art Open Source Software for Speaker Recognition. In *Proceedings of Odyssey : the Speaker and Language Recognition Workshop*, 2008.

[12] L. Breiman. Random Forests. *Machine Learning*, 45(1) :5–32, 2001.

[13] D. Brezeale. Using Closed Captions and Visual Features to Classify Movies by Genre. In *Proceedings of the 7th International Workshop on Multimedia Data Mining*, 2006.

[14] D. Brezeale and D.J. Cook. Automatic Video Classification : A Survey of the Literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews*, 38(3) :416–430, 2008.

[15] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad. Detecting Audio Events for Semantic Video Search. In *InterSpeech*, 2009.

[16] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2) :121–167, 1998.

[17] J.J. Burred. Genetic Motif Discovery Applied to Audio Analysis. In *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[18] R. Cai, L. Lu, H.-J. Zhang, and L.-H. Cai. Highlight Sound Effects Detection in Audio Stream. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 3, pages III – 37–40, 2003.

[19] C. Canton-Ferrer, T. Butko, C. Segura, X. Giro, C. Nadeu, J. Hernando, and J.R. Casas. Audiovisual Event Detection Towards Scene Understanding. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 81–88, 2009.

[20] C.-C. Chang and C.-J. Lin. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2 :27 :1–27 :27, 2011.

[21] E. Charniak. Bayesian Networks Without Tears : Making Bayesian Networks More Accessible to the Probabilistically Unsophisticated. *Artificial Intelligence Magazine*, 12(4) :50–63, 1991.

[22] C.-Y. Chen, A. Abdallah, and W. Wolf. Audiovisual Gunshot Event Recognition. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, volume 6, pages 4807–4812, 2006.

[23] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su. Violence Detection in Movies. In *Proceedings of the 8th International Conference on Computer Graphics, Imaging and Visualization*, pages 119–124, 2011.

[24] L.-H. Chen, C.-W. Su, C.-F. Weng, and H.-Y.M. Liao. Action Scene Detection With Support Vector Machines. *Journal of Multimedia*, 4 :248–253, 2009.

[25] P.-H. Chen, R.-E. Fan, and C.-J. Lin. A Study on SMO-Type Decomposition Methods for Support Vector Machines. *IEEE Transactions on Neural Networks*, 17(4) :893–908, 2006.

[26] Y. Chen, L. Zhang, B. Lin, Y. Xu, and X. Ren. Fighting Detection Based on Optical Flow Context Histogram. In *Innovations in Bio-inspired Computing and Applications (IBICA), 2011 Second International Conference on*, pages 95–98, dec. 2011.

[27] J. Cheng and R. Greiner. Learning Bayesian Belief Network Classifiers : Algorithms and System. In *Proceedings of 14 th Biennial conference of the Canadian Society on Computational Studies of Intelligence : Advances in Artificial Intelligence*, pages 141–151, 2001.

[28] J. Cheng, R.l Greiner, J. Kelly, D. Bell, and W. Liu. Learning Bayesian Networks from Data : an Information-Theory Based Approach. *Artificial Intelligence*, 137 :43–90, 2002.

[29] D. M. Chickering. Learning Equivalence Classes of Bayesian-Network Structures. *Journal of Machine Learning Research*, 2 :445–498, 2002.

[30] M.L. Chin and J.J. Burred. Audio Event Detection Based on Layered Symbolic Sequence Representations. In *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[31] V. Claveau. *Acquisition Automatique de Lexiques Sémantiques pour la Recherche d'Information*. PhD thesis, Université de Rennes 1, December 2003.

[32] C. Clavel, T. Ehrette, and G. Richard. Events Detection for an Audio-Based Surveillance System. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1306–1309, 6-6 2005.

[33] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-Type Emotion Recognition for Future Audio-Based Surveillance Systems. *Speech Communications*, 50(6) :487–503, 2008.

[34] G. F. Cooper and E. Herskovits. A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9 :309–347, 1992.

[35] M. Cristani, M. Bicego, and V. Murino. Audio-Visual Event Recognition in Surveillance Video Sequences. *IEEE Transactions on Multimedia*, 9(2) :257–267, 2007.

[36] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor. On Kernel-Target Alignment. In *Advances in Neural Information Processing Systems 14*, pages 367–373. MIT Press, 2002.

[37] A. Datta, M. Shah, and N. Da Vitoria Lobo. Person-on-Person Violence Detection in Video Data. In *Proceedings of the 16th International Conference on Pattern Recognition*, 2002.

[38] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proceeding of the International Conference on Computer Vision*, volume 3953, pages 288–301. Springer Berlin Heidelberg, 2006.

[39] R. Davis, B. Buchanan, and E. Shortliffe. Production Rules as a Representation for a Knowledge-Based Consultation Program. *Artificial Intelligence*, 8(1) :15–45, 1977.

[40] F.D.M. de Souza, G.C. Chá Andvez, E.A. do Valle, and A. de A Araujo. Violence Detection in Video Using Spatio-Temporal Features. In *Proceedings of the 23rd Conference on Graphics, Patterns and Images*, pages 224 –230, 2010.

[41] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19 :788–798, 2011.

[42] M. Delakis. *Multimodal Tennis Video Structure Analysis with Segment Models*. PhD thesis, Université de Rennes 1, France, October 2006.

[43] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. A Benchmarking Campaign for the Multimodal Detection of Violent Scenes in Movies. In springer, editor, *Proceedings of the ECCV Workshop on Information Fusion in Computer Vision for Concept Recognition*, 2012.

[44] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. The MediaEval 2012 Affect Task : Violent Scenes Detection. In *Proceedings of the MediaEval 2012 Workshop*, volume 927. ceur-ws.org, 2012.

[45] P.Q. Dinh, C. Dorai, and S. Venkatesh. Video Genre Categorization Using Audio Wavelet Coefficients. In *Proceedings of the 5th Asian Conference on Computer Vision*, January 2002.

[46] G. Elidan and N. Friedman. Learning the Dimensionality of Hidden Variables. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 2001.

[47] G. Elidan and N. Friedman. Learning Hidden Variable Networks : The Information Bottleneck Approach. *Journal of Machine Learning Research*, 6 :81–127, 2005.

[48] G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering Hidden Variables : A Structure-Based Approach. In *Neural Information Processing Systems*, pages 479–485. MIT Press, 2001.

[49] S. Essid. *Classification Automatique des Signaux Audio-Fréquences : Reconnaissance des Instruments de Musique*. PhD thesis, Université Pierre et Marie Curie, 2005.

[50] B. Fernando, E. Fromont, D. Muselet, and M. Sebban. Supervised Learning of Gaussian Mixture Models for Visual Vocabulary Generation. *Pattern Recognition*, 45(2) :897–907, 2012.

[51] M. Fradet. *Contribution à la Segmentation de Séquences d'Images au Sens du Mouvement dans un Contexte Semi-Automatique*. PhD thesis, Université de Rennes 1, France, Janvier 2010.

[52] N. Friedman, K. Murphy, and S. Russell. Learning the Structure of Dynamic Probabilistic Networks. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 139–147. Morgan Kaufmann, 1998.

[53] T. Giannakopoulos, D.I. Kosmopoulos, A. Aristidou, and S. Theodoridis. Violence Content Classification Using Audio Features. In *Proceedings of the 4th Helenic Conference on Artificial Intelligence*, pages 502–507, 2006.

[54] T. Giannakopoulos, D.I. Kosmopoulos, A. Aristidou, and S. Theodoridis. A Multi-Class Audio Classification Method With Respect To Violent Content In Movies Using Bayesian Networks. In *Proceedings of the 9th IEEE Workshop on Multimedia Signal Processing*, pages 90–93, 2007.

[55] T. Giannakopoulos, A. Makris, D.I. Kosmopoulos, S. Perantonis, and S. Theodoridis. Audio-Visual Fusion for Detecting Violent Scenes in Videos. In *Artificial Intelligence : Theories, Models and Applications*, Lecture Notes in Computer Science, pages 91–100. Springer Berlin / Heidelberg, 2010.

[56] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao. Detecting Violent Scenes in Movies by Auditory and Visual Cues. In *Proceedings of the 9th Pacific Rim Conference on Multimedia : Advances in Multimedia Information Processing*, pages 317–326, 2008.

[57] G. Gravier, C.-H. Demarty, S. Baghdadi, and P. Gros. Classification-Oriented Structure Learning in Bayesian Networks for Multimodal Event Detection in Videos. *Journal of Multimedia Tools and Applications*, pages 1–17, 2012.

[58] D. Grossman and P. Domingos. Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood. In *Proceedings of the 21st international Conference on Machine learning*, 2004.

[59] N. Haering, R.J. Qian, and M.I. Sezan. A Semantic Event-Detection Approach and its Application to Detecting Hunts in Wildlife Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(6) :857–868, 2000.

[60] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen. Sound Event Detection in Multisource Environments Using Source Separation. In *Workshop on Machine Listening in Multisource Environments*, 2011.

[61] B. Ionescu, J. Schlüter, I. Mironică, and M. Schedl. A Naïve Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2013.

[62] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification. *IEEE Transactions on Multimedia*, 12(6) :523–535, 2010.

[63] R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, D. Li, and J. Louie. A Probabilistic Layered Framework for Integrating Multimedia Content and Context Information. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2057–2060, 2002.

[64] H. Jégou, M. Douze, and C. Schmid. On the Burstiness of Visual Elements. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[65] H. Jégou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 33(1) :117–128, 2011.

[66] Y.-G. Jiang, Q. Dai, C.C. Tan, X. Xue, and C.-W. Ngo. The Shanghai-Hongkong Team at MediaEval2012 : Violent Scene Detection Using Trajectory-based Features. In *Proceedings of the MediaEval 2012 Workshop*, volume 927. ceur-ws.org, 2012.

[67] T. Joachims. Estimating the Generalization Performance of a SVM Efficiently. Technical Report 25, Universität Dortmund, LS VIII, 1999.

[68] T. Joachims. Training Linear SVM in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 2006.

[69] C. Joder, S. Essid, and G. Richard. Temporal Integration for Audio Classification With Application to Musical Instrument Classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1) :174–186, 2009.

[70] K.S. Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1) :11–21, 1972.

[71] P. Kenny. Joint Factor Analysis of Speaker and Session Variability : Theory and Algorithms. Technical report, CRIM, 2006.

[72] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice Modeling with Sparse Training Data. *IEEE Transactions on Speech and Audio Processing*, 13(3) :345 – 354, 2005.

[73] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Factor Analysis Simplified. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 637 – 640, 2005.

[74] P. Kenny and P. Dumouchel. Experiments in Speaker Verification using Factor Analysis Likelihood Ratios. In *Proceedings of Odyssey : the Speaker and Language Recognition Workshop*, pages 219–226, 2004.

[75] E. Kijak. *Structuration Multimodale des Vidéos de Sports par Modèles Stochastiques*. PhD thesis, Univerté de Rennes 1, 2003.

[76] T. Kocka and N.L. Zhang. Dimension Correction for Hierarchical Latent Class Models. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 267–274, 2002.

[77] B. Kriegel. La Violence à la Télévision. Rapport de la Mission d'Évaluation, d'Analyse et de Propositions Relative aux Représentations Violentes à la Télévision. Technical report, Ministère de la Culture et de la Communication, Paris, France, 2003.

[78] E.G. Krug, J.A. Mercy, L.L. Dahlberg, and A.B. Zwi. The World Report on Violence and Health. *The Lancet*, 360(9339) :1083–1088, 2002.

[79] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj. Audio Event Detection From Acoustic Unit Occurence Pattern. In *Proceedings of the 37th International Conference on Accoustics, Speech and Signal Processing*, 2012.

[80] L.I. Kuncheva. *Combining Pattern Classifiers : Methods and Algorithms*. Wiley-Interscience, 2004.

[81] H. Langseth and T.D. Nielsen. Latent Classification Models. *Journal of Machine Learning*, 59 :237–265, 2005.

[82] H. Langseth and T.D. Nielsen. Classification Using Hierarchical Naive Bayes Models. *Journal of Machine Learning*, 63 :135–159, 2006.

[83] I. Laptev. On Space-Time Interest Points. *Journal of Computer Vision*, 64(2-3) :107–123, 2005.

[84] L. Li. A Novel Violent Videos Classification Scheme Based on the Bag of Audio Words Features. In *Proceedings of the 9th International Conference on Information Technology : New Generations*, pages 7–13, 2012.

[85] Stan Z. Li and G.D. Guo. Content-Based Audio Classification and Retrieval Using SVM Learning. In *Proceedings of IEEE International Conference on Multimedia and Expo*, 2000.

[86] T. Li, M. Ogihara, and Q. Li. A Comparative Study on Content-Based Music Genre Classification. In *Proceedings of the 26th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 282–289, 2003.

[87] J. Lin, Y. Sun, and W. Wang. Violence Detection in Movies with Auditory and Visual Cues. In *Proceedings of the International Conference on Computational Intelligence and Security*, pages 561 –565, 2010.

[88] J. Lin and W. Wang. Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training. In *Proceedings of the 10th Pacific-Rim Conference on Multimedia*, pages 930–935, 2009.

[89] Jianhua Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information theory*, 37 :145–151, 1991.

[90] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu. Coherent Bag-of-Audio-Words-Model for Efficient Large-Scale Video Copy Detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 89–96, 2010.

[91] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *Journal of Computer Vision*, 60 :91–110, 2004.

[92] P. Lucas. Restricted Bayesian Network Structure Learning. In *Advances in Bayesian Networks, Studies in Fuzziness and Soft Computing*, pages 217–232, 2002.

[93] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre. A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification. In *Proceedings of Interspeech*, pages 1242–1245, 2007.

[94] D. Matrouf, F. Verdet, M. Rouvier, J.-F. Bonastre, and G. Linarès. Modeling Nuisance Variabilities with Factor Analysis for GMM-based Audio Pattern Classification. *Journal of Computer Speech and Language*, 25(3) :481–498, 2011.

[95] M.F. McKinney and J. Breebaart. Features for Audio and Music Classification. In *Proceeding of the International Society for Music Information Retrieval*, 2003.

[96] S. Moncrieff, C. Dorai, and S. Venkatesh. Affect Computing in Film Through Sound Energy Dynamics. In *Proceedings of the ACM International Conference on Multimedia*, pages 525–527, 2001.

[97] K.P. Murphy. The Bayes Net Toolbox for Matlab. *Journal of Computing Science and Statistics*, 33 :2001, 2001.

[98] P. Naim, P.-H. Wuillemin, P. Leray, O. Pourret, and A. Becker. *Réseaux Bayésiens*. Eyrolles, 2007.

[99] J. Nam, M. Alghoniemy, and A.H. Tewfik. Audio-Visual Content-Based Violent Scene Characterization. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 353–357, 1998.

[100] E.B. Nievas, O.D. Suarez, G.B. García, and R. Sukthankar. Violence Detection in Video Using Computer Vision Techniques. In *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns*, pages 332–339, 2011.

[101] G.T. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M.G. Strintzis. Combining Multimodal and Temporal Contextual Information for Semantic Video Analysis. In *Proceedings of the 16th IEEE Internation Conference on Image Processing*, pages 4325–4328, 2009.

[102] G. Pass, R. Zabih, and J. Miller. Comparing Images Using Color Coherence Vectors. In *Proceedings of the 4th ACM International Conference on Multimedia*, pages 65–73, 1996.

[103] J. Pearl. *Causality : Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[104] T. Pellegrini, J. Portelo, I. Trancoso, A. Abad, and M. Bugalho. Hierarchical Clustering Experiments for Application to Audio Event Detection. In *Proceedings of the 13th International Conference on Speech and Computer*, 2009.

[105] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. De la Détection d'Évènements Sonores Violents par SVM dans les Films. In *Actes du Congrès des jeunes chercheurs en vision par ordinateur*, 2011.

[106] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies. In *Proceedings of the 37th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[107] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Audio Event Detection in Movies using Multiple Audio Words and Contextual Bayesian Networks. In *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, 2013.

[108] T. Perperis, T. Giannakopoulos, A. Makris, D.I. Kosmopoulos, S. Tsekeridou, S.J. Perantonis, and S. Theodoridis. Multimodal and Ontology-Based Fusion Approaches of Audio and Visual Processing for Violence Detection in Movies. *Journal of Expert Systems with Applications*, 38(11) :14102–14116, 2011.

[109] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis. Gunshot Detection in Audio Streams from Movies by Means of Dynamic Programming and Bayesian Networks. In *Proceedings of the 33rd IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 21–24, 2008.

[110] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro. Non-Speech Audio Event Detection. In *Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1973 –1976, 2009.

[111] M. Ramona and G. Richard. Segmentation Parole/Musique Par Machines à Vecteurs de Support. In *Actes des Journées d'Etude de la Parole*, 2008.

[112] M. Ramona and G. Richard. Comparison Of Different Strategies For A SVM-Based Audio Segmentation. In *Proceedings of the European Conference on Signal Processing*, 2009.

[113] E. Ravelli, G. Richard, and L. Daudet. Audio Signal Representations for Indexing in the Transform Domain. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :434 –446, march 2010.

[114] G. Richard, M. Ramona, and S. Essid. Combined Supervised and Unsupervised Approaches for Automatic Segmentation of Radiophonic Audio Streams. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages II–46–464, 2007.

[115] M. Rouvier, D. Matrouf, and G. Linarès. Factor Analysis for Audio-Basedased Video Genre Classification. In *Proceedings of Interspeech*, 2009.

[116] J. Saunders. Real-Time Discrimination of Broadcast Speech/Music. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 993–996, 1996.

[117] B. Schiele and J.L. Crowley. Object Recognition Using Multidimensional Receptive Field Histograms. In *Proceedings of the European Conference in Computer Vision*, pages 610–619. Springer, 1996.

[118] J. Schlüter, B. Ionescu, I. Mironică, and M. Schedl. ARF @ MediaEval 2012 : An Uninformed Approach to Violence Detection in Hollywood Movies . In *Proceedings of the MediaEval 2012 Workshop*, volume 927. ceur-ws.org, 2012.

[119] C.G.M. Snoek and M. Worring. Multimodal Video Indexing : A Review of the State-of-the-Art. *Journal of Multimedia Tools and Applications*, 25 :5–35, 2003.

[120] M. Soleymani, M. Pantic, and T. Pun. Multi-Modal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing*, 1 :211–223, 2012.

[121] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, second edition, 2001.

[122] J. Sun, X. Wu, S. Yan, L. Fah C., T.-S. Chua, and J. Li. Hierarchical Spatio-Temporal Context Modeling for Action Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2004–2011, 2009.

[123] I. Trancoso, T. Pellegrini, J. Portelo, H. Meinedo, M. Bugalho, A. Abad, and J. Neto. Audio Contributions to Semantic Video Search. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 630–633, 2009.

[124] I. Trancoso, J. Portelo, M. Bugalho, J.P. Neto, and A.J. Serralheiro. Training Audio Events Detectors with a Sound Effects Corpus. In *Proceedings of Interspeech*, pages 2546–2549, 2008.

[125] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface. Channel Factors Compensation in Model and Feature Domain for Speaker Recognition. In *Proceedings of Odyssey : the Speaker and Language Recognition Workshop*, pages 1–6, 2006.

[126] F. Vallet, S. Essid, J. Carrive, and G. Richard. Robust Visual Features for the Multimodal Identification of Unregistered Speaker in TV Talk-Shows. In *Proceedings of the IEEE Internation Conference on Image Processing*, 2010.

[127] N. Vasconcelos and A. Lippman. Towards Semantically Meaningful Feature Spaces for the Characterization of Video Content. In *Proceedings of the IEEE International Conference on Image Processing*, volume 1, pages 25–28, 1997.

[128] J. Vendrig and M. Worring. Interactive Adaptive Movie Annotation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 1, pages 93–96, 2002.

[129] R. Vogt and S Sridharan. Explicit Modelling of Session Variability for Speaker Verification. *Journal of Computer Speech and Language*, 22 :17–38, 2008.

[130] S. Wang, S. Jiang, Q. Huang, and W. Gao. Shot Classification for Action Movies Based on Motion Characteristics. In *Proceedings of the IEEE International Conference on Image Processing*, pages 2508–2511, 2008.

[131] Y. Wang, Z. Liu, and J.-C. Huang. Multimedia Content Analysis - Using Both Audio and Visual Clues. *Signal Processing Magazine*, 17(6) :12–36, 2000.

[132] U. Westermann and R. Jain. Toward a Common Event Model for Multimedia Applications. *Journal of Multimedia*, 14(1) :19–29, 2007.

[133] Y. Wu, E.Y. Chang, K.C.-C. Chang, and J.R. Smith. Optimal Multimodal Fusion for Multimedia Data Analysis. In *Proceedings of the 12th ACM International Conference on Multimedia*, pages 572–579, 2004.

[134] L. Xie, H. Sundaram, and M. Campbell. Event Mining in Multimedia Streams. *Proceedings of the IEEE*, 96(4) :623–647, 2008.

[135] J. Yang, Y.-G. Jiang, A.G. Hauptmann, and C.-W. Ngo. Evaluating Bag-of-Visual-Words Representations in Scene Classification. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pages 197–206, 2007.

[136] A. Yoshitaka and M. Miyake. Scene Detection by Audio-Visual Features. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 48–51, 2001.

[137] N.L. Zhang. Hierarchical Latent Class Models for Cluster Analysis. In *Proceedings of the 18th National Conference on Artificial intelligence*, pages 230–237, 2002.

[138] N.L. Zhang, T.D. Nielsen, and F.V. Jensen. Latent Variable Discovery in Classification Models. *Journal of Artificial Intelligence in Medicine*, 3 :283–299, 2004.

[139] X. Zou, O. Wu, Q. Wang, W. Hu, and J. Yang. Multi-modal Based Violent Movies Detection in Video Sharing Sites. In *Intelligent Science and Intelligent Data Engineering*, volume 7751 of *Lecture Notes in Computer Science*, pages 347–355. Springer Berlin Heidelberg, 2013.

Liste de publications

Conférences internationales

[1] J. Fleureau, C. Penet, P. Guillotel, and C.-H. Demarty. Electrodermal Activity Applied to Violent Scenes Impact Measurement and User Profiling. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012.

[2] C. Penet, C.-H. Demarty, G. Gravier, and P Gros. Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies. In *ICASSP - 37th International Conference on Acoustics, Speech, and Signal Processing (2012)*, Kyoto, Japon, March 2012.

[3] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Audio Event Detection using Multiple Audio Words and Contextual Network. In *CBMI - 11th International Workshop on Content-Based Multimedia Indexing*, 2013. **[Best Paper Award]**.

Workshops internationaux

[1] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. L. Quang, and Y.-G. Jiang. The MediaEval 2013 Affect Task : Violent Scenes Detection. In *MediaEval 2013 Workshop*, 2013.

[2] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. The MediaEval 2011 Affect Task : Violent Scenes Detection in Hollywood Movies. In *MediaEval 2011 Workshop*, volume 807, Pisa, Italy, September 1-2 2011. ceur-ws.org.

[3] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. A Benchmarking Campaign for the Multimodal Detection of Violent Scenes in Movies. In springer, editor, *ECCV 2012 Workshop on Information Fusion in Computer Vision for Concept Recognition*, 2012.

[4] Claire-Hélène Demarty, Cédric Penet, Guillaume Gravier, and Mohammad Soleymani. The MediaEval 2012 Affect Task : Violent Scenes Detection. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012. ceur-ws.org.

[5] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Technicolor and INRIA/IRISA at MediaEval 2011 : learning temporal modality integration with Bayesian Networks. In *MediaEval 2011 Workshop*, volume 807, Pisa, Italy, September 1-2 2011. ceur-ws.org.

[6] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. Technicolor/INRIA Team at the MediaEval 2013 Violent Scenes Detection Task. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 2013.

[7] C. Penet, C.-H. Demarty, M. Soleymani, G. Gravier, and P. Gros. Technicolor/INRIA/Imperial College London at the MediaEval 2012 Violent Scene Detection Task. In *MediaEval 2012 Workshop*, volume 927, Pisa, Italy, October 4-5 2012. ceur-ws.org.

Conférences nationales

[1] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. De la détection d'évènements sonores violents par SVM dans les films. In *ORASIS - Congrès des jeunes chercheurs en vision par ordinateur*, Praz-sur-Arly, France, 2011. INRIA Grenoble Rhône-Alpes.

Brevets

[1] F. Botta, P. Gallardo, C. Penet, and C.-H. Demarty. Method for setting a watching level for an audiovisual content, 2013.

[2] M. Fradet, A. Newson, and C. Penet. Method for processing an audiovisual content and corresponding device, 2013.

Table des figures

Liste des tableaux

Résumé

Dans cette thèse, nous nous intéressons à la détection de concepts sémantiques dans des films "Hollywoodiens" à l'aide de concepts audio et vidéos, dans le cadre applicatif de la détection de violence. Nos travaux se portent sur deux axes : la détection de concepts audio violents, tels que les coups de feu et les explosions, puis la détection de violence, dans un premier temps uniquement fondée sur l'audio, et dans un deuxième temps fondée sur l'audio et la vidéo.

Dans le cadre de la détection de concepts audio, nous mettons tout d'abord un problème de généralisation en lumière, et nous montrons que ce problème est probablement dû à une divergence statistique entre les attributs audio extraits des films. Nous proposons pour résoudre ce problème d'utiliser le concept des mots audio, de façon à réduire cette variabilité en groupant les échantillons par similarité, associé à des réseaux Bayésiens contextuels. Les résultats obtenus sont très encourageants, et une comparaison avec un état de l'art obtenu sur les même données montre que les résultats sont équivalents. Le système obtenu peut être soit très robuste vis-à-vis du seuil appliqué en utilisant la fusion précoce des attributs, soit proposer une grande variété de points de fonctionnement. Nous proposons enfin une adaptation de l'analyse factorielle développée dans le cadre de la reconnaissance du locuteur, et montrons que son intégration dans notre système améliore les résultats obtenus.

Dans le cadre de la détection de violence, nous présentons la campagne d'évaluation MediaEval Affect Task 2012, dont l'objectif est de regrouper les équipes travaillant sur le sujet de la détection de violence. Nous proposons ensuite trois systèmes pour détecter la violence, deux fondés uniquement sur l'audio, le premier utilisant une description TF-IDF, et le second étant une intégration du système de détection de concepts audio dans le cadre de la détection violence, et un système multimodal utilisant l'apprentissage de structures de graphe dans des réseaux bayésiens. Les performances obtenues dans le cadre des différents systèmes, et une comparaison avec les systèmes développés dans le cadre de MediaEval, montrent que nous sommes au niveau de l'état de l'art, et révèlent la complexité de tels systèmes.

Abstract

In this thesis, we focus on the detection of semantic concepts in "Hollywood" movies using audio and video concepts for the detection of violence. We present experiments in two main areas : the detection of violent audio concepts such as gunshots and explosions, and the detection of violence, initially based only on audio, then based on both audio and video.

In the context of audio concepts detection, we first show a generalisation arising between movies. We show that this problem is probably due to a statistical divergence between the audio features extracted from the movies. In order to solve it, we propose to use the concept of audio words, so as to reduce the variability by grouping samples by similarity, combined with contextual Bayesian networks. The results are very encouraging, and a comparison with the state of the art obtained on the same data shows that the results we obtain are equivalent. The resulting system can be either robust against the threshold applied by using early fusion of features, or provides a wide variety of operating points. We finally propose an adaptation of the factor analysis scheme developed in the context of speaker recognition, and show that its integration into our system improves the results.

In the context of the detection of violence, we present the Mediaeval Affect Task 2012 evaluation campaign, which aims at bringing together teams working on the topic of violence detection. We then propose three systems for detecting the violence. The first two are based only on audio, the first using a TF-IDF description, and the second being the integration of the previous system for the detection violence. The last system we present is a multimodal system based on Bayesian networks that allows us to explore structure learning algorithms for graphs. The performance obtained in the different systems, and a comparison to the systems developed within Mediaeval, show that we are comparable to the state of the art, and show the complexity of such systems.