N$^{\text{O}}$ d'ordre : 2464

THÈSE

présentée devant

l'Université de Rennes 1

pour obtenir

le grade de :	Docteur de l'Université de Rennes 1
Mention :	Traitement du Signal et Télécommunications

par

Mouhamadou SECK

Équipe d'accueil :	IRISA, Projet Sigma2
École Doctorale :	Mathématiques, Informatique, Signal et Électronique et Télécommunications
Composante Universitaire :	Structures et Propriétés de la Matière

Titre de la thèse :

DÉTECTION DE RUPTURES ET SUIVI DE CLASSE DE SONS POUR L'INDEXATION SONORE

Soutenue le 5 Janvier 2001 devant la commission d'examen :

Mr	Gérard	Chollet	Président
Mme	Régine	André-Obrecht	Rapporteur
Mr	Jean-François	Bonastre	Rapporteur
Mr	Christian	Wellekens	Examinateur
Mr	Bernard	Delyon	Directeur
Mr	Frédéric	Bimbot	Co-Directeur

À mes très chers parents et à toute ma famille pour toutes ces années passées loin d'eux.

À Mikaëla, pour m'avoir soutenu au quotidien et surtout pour sa patience et sa compréhension.

Remerciements

Mes premiers remerciements s'adressent à Frédéric BIMBOT et Bernard DELYON pour avoir bien voulu diriger ces travaux.

Je témoigne ma reconnaissance à Bernard DELYON pour avoir encadré mes premiers travaux de recherche à l'IRISA, dans le cadre de mon stage de D.E.A. C'est dans ce contexte que j'ai fait la rencontre de Frédéric BIMBOT avec qui j'ai pris beaucoup de plaisir à travailler.

Je tiens à remercier très chaleureusement Frédéric pour la qualité de son endacrement, aussi bien sur le plan scientifique que sur la dimension humaine. De par ses orientations scientifiques, son écoute, sa disponibilité et son soutien au quotidien, Frédéric a joué un rôle très important dans l'accomplissement de ces travaux.

Pour avoir significativement contribué à un très agréable cadre de réalisation de cette thèse, j'adresse mes remerciements à tous les membres du Groupe Parole et Sons [1] du projet Sigma2, plus particulièrement Raphaël BLOUET avec qui j'ai plus étroitement collaboré, mais aussi Laurent BENAROYA, Bruno JACOB, Ivan MAGRIN-CHAGNOLLEAU, Lorcan MC DONAGH. Ces remerciements s'adressent également à toute l'équipe Sigma2, notamment les doctorants Jean-Luc LECALVEZ, Nicolas MOAL, Rodolphe PRIAM et Laurent TROMP.

Je voudrais vivement remercier les personnels technique, administratif et du centre de documentation de l'IRISA qui, de par leur disponibilité, ont énormément contribué aux

1. Actuel projet METISS

appréciables conditions dans lesquelles ce travail a été effectué.

J'exprime toute ma gratitude aux rapporteurs de cette thèse, Jean-François BONASTRE et Régine ANDRÉ-OBRECHT, pour l'attention qu'ils ont accordé à ces travaux. Je remercie aussi Gérard CHOLLET pour avoir bien voulu présider le jury de soutenance de cette thèse, ainsi que Christian WELLEKENS, membre du jury en qualité d'examinateur, pour l'intérêt qu'il a bien voulu témoigner à mes travaux, à travers les échanges fructueux que nous avons eus.

Présentation

Mots-clés : indexation, signaux sonores, tests statistiques, détection de ruptures, modèles de mélange, détection et suivi de classe de sons.

Le développement rapide que connaissent aujourd'hui les nouvelles technologies de l'information a contribué à une production massive de documents multimédia. Dans ce contexte, le besoin en outils de gestion automatique de bases de données sonores va en croissant, et ne s'est jamais autant fait ressentir.

L'objectif de cette thèse est de définir, mettre en œuvre, étudier et évaluer des méthodes automatiques d'analyse de documents sonores, dans le but d'extraire des descripteurs pour l'indexation. Compte tenu de l'état actuel des connaissances dans le domaine de l'indexation sonore, il serait ambitieux d'envisager une description permettant de caractériser de manière complète un document sonore. Ce travail de thèse se concentre donc sur trois types de descripteurs, en réponse à trois problèmes :

1. La détection de changements de plage sonore;
2. La détection d'une classe de sons dans un document;
3. Le suivi d'une classe de sons dans un document.

Le cadre dans lequel est placée la détection de changements de plage sonore, est celui de la détection de ruptures statistiques sur un signal aléatoire à temps discret. Les approches auxquelles on s'intéresse dans cette thèse, consistent d'abord à calculer à chaque instant un

indice de présence de rupture, puis à extraire à partir de celui-ci un critère de décision de rupture. Les indices utilisés sont dérivés de statistiques de test d'existence d'une rupture aux différents instants. Et le critère est plus précisément obtenu par mise en évidence des extrema dominants de l'indice. La décision finale de présence d'une rupture à un instant donné est prise en comparant la valeur du critère à un seuil.

Pour le test d'existence d'une rupture, la statistique la plus utilisée est le rapport de vraisemblance généralisé. Cependant elle est difficile à interpréter sous l'hypothèse de présence de rupture. Pour cette raison, une statistique alternative, basée sur une estimation de la divergence de Kullback entre les distributions avant et après rupture, est proposée. D'autres statistiques mesurant la même quantité existent déjà, mais dans un cadre de détection séquentielle de ruptures. L'étude de la loi asymptotique de cette statistique alternative, sous l'hypothèse d'absence de rupture, met en évidence beaucoup de ressemblance avec la loi asymptotique de la statistique usuelle du rapport de vraisemblance généralisé.

Nous proposons également une méthode efficace de mise en évidence des extrema dominants d'une fonction à temps discret, en vue de l'utiliser comme méthode d'extraction du critère de rupture à partir d'un indice. L'avantage majeur par rapport à d'autres méthodes pré-existantes, est de n'utiliser aucun seuil intermédiaire, permettant ainsi de s'affranchir des difficultés de réglage d'un tel paramètre pour l'adaptation (souvent nécessaire) aux types de signaux et de ruptures traités.

La détection d'une classe de sons (cible) dans un document est formulée comme un problème de test statistique entre deux hypothèses : l'absence et la présence du son cible dans le document. Le test du rapport de vraisemblance classique entre les distributions (des vecteurs acoustiques) cible et non-cible est théoriquement inadapté dans le cas d'un document où le son cible peut être mélangé à d'autres sons non-cibles.

Pour cette raison, deux autres statistiques de test sont étudiées. Elles sont toutes les deux basées sur une modélisation de la distribution (des vecteurs acoustiques extraits) du document, par un mélange de deux distributions: la distribution cible et la distribution non-cible. Les deux statistiques dérivées de ce modèle de mélange sont : l'estimateur du maximum de vraisemblance du poids de la distribution cible (calculé à partir du document observé), et le rapport de vraisemblance généralisé du test d'absence versus présence du son cible.

Le rapport de vraisemblance généralisé est la seule statistique dont la loi asymptotique en absence de la classe de sons cible, est indépendante des distributions cible et non-cible. Évaluée sur une tâche de détection du locuteur dans une conversation téléphonique (corpus Switchboard), le poids de la distribution cible conduit à de meilleures performances que le rapport de vraisemblance usuel.

Les algorithmes de suivi d'une classe de sons utilisent conjointement les méthodes de détection d'une classe de sons et celles de détection de changements de plage sonore. Nous avons testé les différentes approches étudiées sur des bases de données réelles et d'un grand volume. Il s'agit plus précisément d'expériences de suivi de parole sur un documentaire (corpus AIM de l'INA, projet DiVAN) et de suivi du locuteur dans une conversation téléphonique (corpus Switchboard, évaluations NIST au sein du consortium ELISA). Sur la tâche de suivi de parole, les trois statistiques de test donnent des niveaux de performances sensiblement comparables. Cependant, en concordance avec la théorie, le poids de la distribution cible fournit de meilleures performances que le rapport de vraisemblance usuel, sur la tâche de suivi du locuteur.

Table des matières

Table des figures

Introduction générale

Les progrès effectués ces dernières années dans le domaine des nouvelles technologies de l'information ont fortement contribué à la production massive de documents multimédia. Ce développement rapide des nouvelles technologies a, par la même occasion, rendu plus facile l'accès grand-public et la circulation de documents. Toutefois, la gestion de ce flux de données devient de moins en moins aisée car les outils de gestion ne connaissent pas un essor aussi rapide que la production de documents. En réponse à ce besoin, on assiste actuellement à un regain d'intérêt pour la gestion automatique de bases de documents multimédia, de la part de la communauté scientifique, soutenu par les besoins exprimés par les acteurs publics et privés. Ainsi, au niveau national et international, de nouveaux groupes de travail, conférences et projets, portant sur l'indexation de documents multimédia par le contenu, voient le jour.

La gestion d'une base de documents multimédia tourne autour des trois points suivants :

- l'extraction d'information à partir d'un document,
- la représentation des informations extraites,
- la recherche d'informations en réponse à une requête dans la base.

Les deux premiers points correspondent à l'indexation de documents. L'extraction d'informations peut être définie comme une opération qui consiste à associer des attributs quantitatifs ou qualificatifs (souvent cognitifs), à un document. Ce point repose donc sur la définition d'un lexique de descripteurs (language d'indexation) d'une part, et d'un processus d'extraction, d'autre part. Par exemple, il peut s'agir de qualifier un document sonore comme étant de la musique, de la parole, la voix d'un individu donné, etc...

Le deuxième point a pour but de représenter les informations extraites des documents, relativement à une optique bien fixée. D'usage, on distingue deux types d'indexation : l'indexation orientée requête et l'indexation orientée document [Berrut, 1997]. La première est effectuée dans le souci de couvrir, autant que possible, le champ des requêtes de recherche auxquelles le document peut répondre. Quant à l'indexation orientée document, elle cherche à résumer le document par un ensemble d'informations pertinentes, en vue d'une caractérisation (en toute généralité) du document. L'indexation orientée requête induit alors une organisation de la base de documents, dans le but de faciliter les processus de recherche d'information. Par exemple, les documents peuvent être agencés de manière à rendre plus aisée la navigation dans la base de données.

L'indexation automatique de documents multimédia est un domaine relativement récent, où les travaux actuels portent encore sur la recherche de définitions concises et unifiées des différentes problématiques et schémas d'indexation. Sur ces points, la vidéo connaît plus d'avancées que l'audio. Le présent travail concerne exclusivement le contenu audio du document.

L'objet de cette thèse est de définir, mettre en œuvre, étudier et évaluer des algorithmes automatiques d'analyse de documents sonores, dans le but d'extraire des descripteurs pour l'indexation.

L'ambition de ce travail de thèse n'est naturellement pas de couvrir l'ensemble des descripteurs permettant une caractérisation exhaustive d'un document sonore. Le choix des types de descripteurs sur lesquels porte ce travail est guidé par la volonté de servir aussi bien à une caractérisation du document par rapport à d'autres documents, qu'à une structuration interne du document. Les trois familles de descripteurs auxquelles on s'intéresse ici, répondent aux trois problèmes présentés succinctement ci-après.

1. **La détection de changements de plage sonore**. C'est une tâche de segmentation

qui peut être utilisée comme phase de prétraitement dans certaines applications en Traitement de la Parole. Elle peut également servir à indiquer une organisation temporelle du document en plages sonores homogènes. Le descripteur en réponse à ce problème est une liste d'instants correspondant à des changements de plages;

2. **La détection d'une classe de sons dans un document**. Il s'agit simplement ici de déterminer si un type de son fixé (dit cible) est présent ou non dans un document. Le descripteur est une réponse binaire indiquant la présence ou non du type de son cible. Comme exemple d'application, citons la vérification du locuteur qui consiste à dire si un individu donné est présent dans un document sonore;

3. **Le suivi d'une classe de sons dans un document**. Un exemple classique d'application est la détermination des plages où est présente de la parole ou de la musique dans un document (comme une émission de variétés).

Le travail présenté dans cette thèse s'appuie sur des méthodes issues de trois branches des Mathématiques Appliquées et étroitement liées entre elles : le Traitement du Signal, les Probabilités et les Statistiques. Pour chacun des algorithmes d'analyse du document sonore étudié, le signal sonore est d'abord paramétrisé en un ensemble de vecteurs acoustiques, en faisant appel à des méthodes de Traitement du Signal. Cette phase de paramétrisation est suivie d'un traitement basé sur des méthodes probabilistes et statistiques, dépendant de la tâche à accomplir. Un des inconvénients de ces méthodes est le coût de calcul. Mais avec les puissances de calcul des processeurs actuels, et qui vont en croissant très rapidement, ces méthodes gagnent en applicabilité.

Cette thèse a été effectuée en interaction avec le projet européen Esprit DiVAN, portant sur l'indexation de données audiovisuelles. L'indexation est un domaine qui réunit divers acteurs. La collaboration de tous ces acteurs est un facteur nécessaire à l'essor d'un domaine. C'est dans cette optique que s'inscrit le projet DiVAN qui rassemble des scientifiques [2],

2. ICS-FORTH (Grèce), IRISA (France), et Technical University of Darmstadt (Allemagne).

des industriels[3] et des utilisateurs[4] pour la conception, l'évaluation, le développement et la commercialisation d'outils de gestion de base de données audiovisuels. De le cadre de ce projet, ce travail a été effectué au sein de l'équipe SIGMA2 de l'IRISA, en coordination avec une autre thèse en image au sein de l'équipe VISTA de l'IRISA, sur l'analyse des mouvements dans les scènes vidéo.

La plan de ce travail s'articule en trois parties :

La première partie présente les notions générales de Traitement du Signal et de Statistiques, utilisées dans les autres parties. Elle définit l'architecture générale des systèmes d'extraction de descripteurs étudiés et décrit les différents modules qui la constituent. Elle rappelle également quelques concepts de base de la théorie statistique des tests d'hypothèses.

La seconde partie est consacrée à la détection de changements de plage sonore. Les méthodes qui y sont abordées sont basées sur la détection de ruptures statistiques sur un signal à temps discret. Le principe de la méthode de détection de ruptures proposée consiste à calculer d'abord à chaque instant un indice de rupture, à partir duquel est extrait ensuite un critère de décision. Les indices utilisés correspondent à des statistiques de test d'existence d'une rupture en un instant donné.

Dans ce contexte, deux statistiques de détection d'une rupture sont étudiées et comparées : d'une part le rapport de vraisemblance généralisé usuel, et d'autre part une statistique alternative basée sur une mesure de la divergence de Kullback[5] entre les distributions avant et après rupture. Des propriétés de cette statistique alternative ont été établies pour des familles de distributions de type exponentiel (Gaussien, Exponentiel, Poisson, etc..). La loi asymptotique de cette statistique alternative, sous l'hypothèse d'absence de rupture, est

3. INTRACOM et TECMATH (Allemagne).
4. INA (France), RAI (Italie) et ERT (Grèce).
5. D'autres statistiques de détection d'une rupture, basées sur la divergence de Kullback de rupture ont déjà été proposées dans la littérature, mais dans un cadre de détection séquentielle de ruptures.

également énoncée, et présente beaucoup de similarités avec la loi asymptotique (connue) en absence de rupture, de la statistique du rapport de vraisemblance généralisé (pour la détection d'une rupture).

Le critère est obtenu en appliquant à l'indice une méthode nouvelle d'extraction des extrema dominants d'une fonction à temps discret. L'atout majeur de cette méthode, par rapport à d'autres méthodes existantes, est de n'utiliser aucun seuil intermédaire.

La troisième partie traite de la détection et du suivi d'une classe de sons (dite cible), dans un document sonore. Dans cette partie, les différents problèmes abordés sont placés dans le cadre des tests statistiques d'hypothèses. Trois statistiques de test ont été l'objet de notre étude : le rapport de vraisemblance usuel entre la distribution (des vecteurs acoustiques) cible et celle non-cible, et deux autres statistiques introduites dans ce travail, qui sont basées sur une modélisation des vecteurs acoustiques du document par un mélange de deux distributions : la distribution cible et la distribution non-cible. Les deux statistiques dérivées du modèle de mélange sont alors : l'estimateur du maximum de vraisemblance du poids de la distribution cible, basé sur les vecteurs acoustiques du document, et le rapport de vraisemblance généralisé du test d'absence versus présence du son cible. L'étude des lois asymptotiques des trois statistiques de test suggère une préférence pour le rapport de vraisemblance généralisé, car elle est la seule dont la loi asymptotique est indépendante des deux distributions cible et non-cible.

Les algorithmes présentés dans cette troisième partie ont été évalués dans le cadre des campagnes annuelles d'évaluations NIST (National Institut of Standards and Technologies), portant sur le thème de la Reconnaissance Automatique du Locuteur, et auxquelles l'équipe SIGMA2 de l'IRISA participe depuis trois ans (1998, 1999 et 2000) au sein du Consortium ELISA, réunissant des laboratoires pour la plupart européens[6] qui partagent des ressources logicielles et mettent en œuvre une plate-forme commune dans le but de faciliter leur

6. Le consortium ELISA est fondé en 1997 par l'ENST (France), l'EPFL (Suisse), l'IDIAP (Suisse), l'IRISA (France), le LIA (France). Ils ont été rejoints en 1998 par le VUT (République Tchèque), le RMA (Belgique), le RICE University (USA), et en 1999 par le FPMS (Belgique).

participation aux campagnes d'évaluations NIST. Les résultats présentés concernent les évaluations 1999 et 2000. En 1999, nos systèmes se sont relativement bien comportés par rapport aux autres systèmes. Cependant, en 2000, nos systèmes ont été distancés en termes de performances par les meilleurs systèmes soumis aux évaluations NIST. Des travaux complémentaires laissent paraître des perspectives d'amélioration de nos systèmes en vue des évaluations NIST 2001. D'abord en s'inspirant de la paramétrisation acoustique et de la modélisation des meilleurs systèmes, mais aussi en utilisant la statistique du rapport de vraisemblance généralisé.

La structure de ce document reflète les différentes composantes du travail effectué. Après une première partie qui rappelle les principaux outils théoriques utilisés dans ces travaux, une seconde partie est consacrée à la détection de changements de plage sonore, et une troisième partie à la détection et au suivi d'une classe de sons dans un document sonore, accompagnée d'évaluations sur un grand volume de données réelles.

Première partie

GÉNÉRALITÉS

Présentation

L'objet de cette partie est de présenter, sous une architecture commune, les différents systèmes d'extraction de descripteurs ainsi que les outils théoriques de Traitement du Signal, de Probabilités et de Statistiques utilisés par les différents systèmes.

Le premier chapitre donne une vue de l'ensemble de l'architecture des systèmes d'extraction de descripteurs et décrit plus en détail les différents modules qui la compose. Ce chapitre permet notamment de préciser les outils théoriques qui interviennent dans la conception des différents traitements réalisés.

Le second chapitre rappelle des concepts de la théorie statistique des tests, cadre dans lequel sont placés les modules de calcul de score et de décision. Ce chapitre pose les aspects théoriques qui permettent de mieux prévoir et comprendre les performances des modules de calcul du score et de décision.

Chapitre 1

Structure générale des systèmes d'extraction de descripteurs

Les systèmes étudiés suivent une architecture classique, qui est composée de quatre modules :

- un module de paramétrisation acoustique,
- un module de modélisation,
- un module de calcul de score,
- un module de décision.

La figure 1.1 schématise l'architecture des systèmes. Dans cette figure, les boîtes ovales correspondent aux différents modules, et les boîtes rectangulaires en pointillés sont les objets échangés par les modules. Les flèches en traits pointillés désignent les entrées vers les modules, tandis que les flèches en traits pleins indiquent les sorties des modules.

Le signal sonore est systématiquement paramétrisé. La paramétrisation consiste à extraire du signal un ensemble de vecteurs paramètres acoustiques. Elle commence d'abord par définir régulièrement sur le signal, des intervalles de temps d'égale durée, que l'on appelle usuellement trames. Il est d'usage de considérer une certaine durée de recouvrement entre trames successives. Chaque trame est ensuite représentée par un certain nombre de paramètres, constituant un vecteur acoustique. La section 1.1 présente de manière non exhaustive les

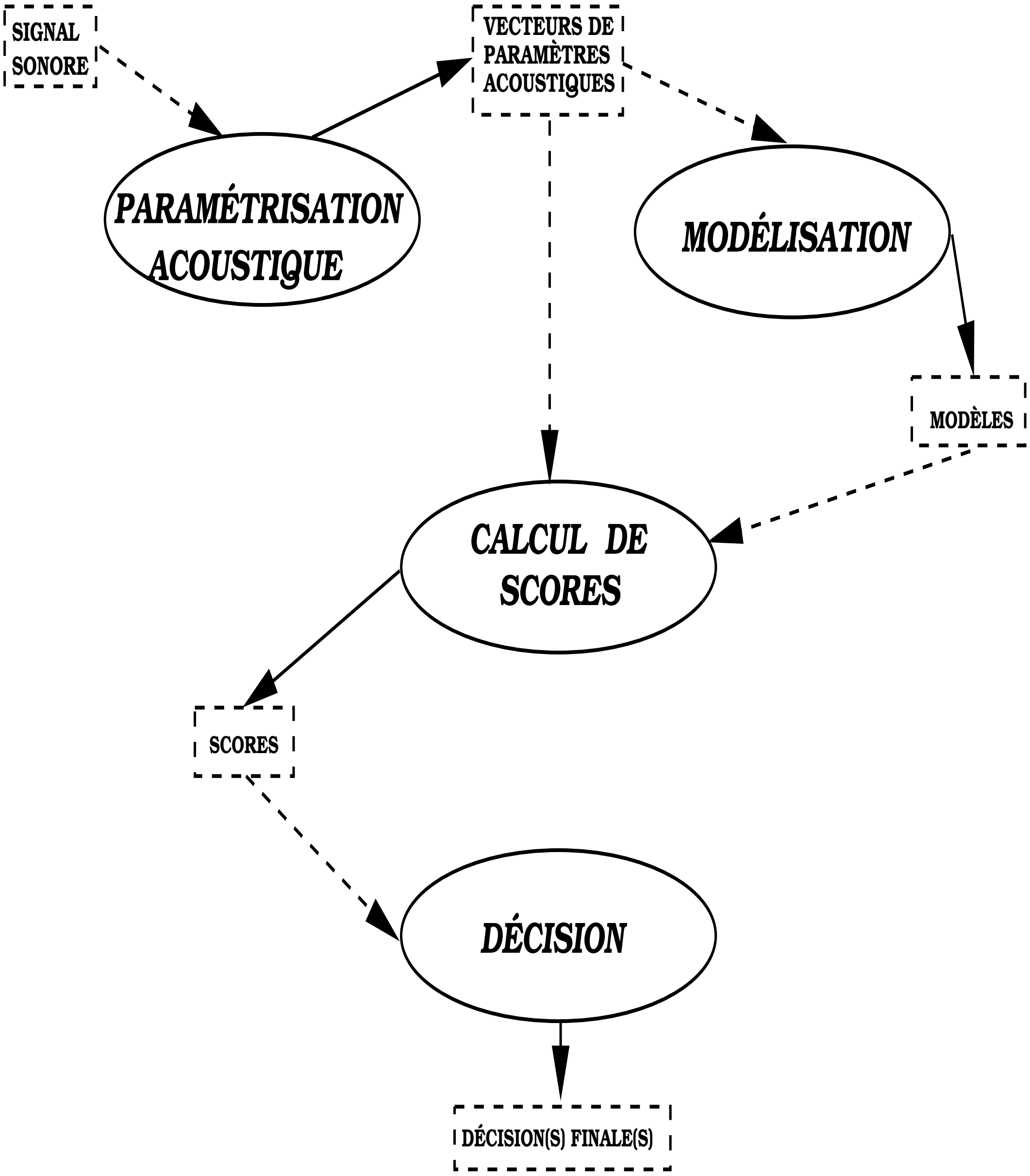

FIG. 1.1 – *Architecture générale des systèmes d'extraction de descripteurs.*

paramètres acoustiques les plus utilisés en Traitement Automatique du Signal Sonore.

Pour mieux fixer les idées, la description des autres modules va s'appuyer sur un exemple de tâche, qui est la vérification du locuteur. Cette tâche consiste à dire si un locuteur X est présent dans un document sonore.

Le module de modélisation a pour fonction de modéliser la distribution (de probabilité) d'un ensemble de vecteurs acoustiques. Cet ensemble de vecteurs est, dans la pratique, extrait de signaux d'apprentissage appartenant à une même classe de sons comme la musique, la parole, la parole d'un locuteur, etc... Le modèle ainsi obtenu représente d'une certaine manière la classe de sons. Par exemple, pour la vérification du locuteur, deux modèles sont appris : un modèle du locuteur cible noté X et un modèle plus complexe qui représente tout ce qui n'est pas le locuteur cible, dénommé $\overline{X}$. Dans ce travail, une famille paramétrique de modèles est utilisée : les Modèles de Mélanges de Gaussiennes (MMG).

Le module de calcul des scores vise à définir des fonctions (scores) des vecteurs acoustiques extraits d'un signal de test, qui guident la décision. Ces fonctions sont généralement définies à l'aide de modèles préalablement appris. En revenant sur l'exemple de la vérification du locuteur, le score est défini par les modèles X et $\overline{X}$.

Les scores sont ensuite transmis au module de décision qui prend la décison finale (par exemple : acceptation ou rejet de l'identité proclamée en vérification du locuteur). Le module de calcul des scores et celui de décision doivent être conçus en adéquation l'un avec l'autre. L'approche générale classique utilisée en vérification du locuteur (généralement dans un système qui fournit une réponse binaire) consiste à définir un score d'autant plus grand que la voix du locuteur cible X est proche de celui présent dans le document, puis à prendre la décision par comparaison à un seuil.

1.1 Paramétrisation acoustique

La paramétrisation acoustique est la phase initiale et systématique de tous les systèmes d'extraction de descripteurs étudiés dans ce travail. Une des approches consiste à définir la paramétrisation acoustique en fonction du type de problème abordé, comme dans [Scheirer et Slanley, 1997, Saunders, 1996] pour la discrimination parole/musique. L'inconvénient majeur de cette approche est qu'elle ne peut pas être transposée comme telle pour traiter d'autres problèmes de discrimination de signaux sonores, bien que ces problèmes puissent être définis sous le même formalisme. Il faudrait donc définir pour chaque problème une paramétrisation optimale. Compte tenu de la diversité des signaux qui peuvent être pris en compte dans certains problèmes, cette approche n'est pas toujours facile à mettre en œuvre. Dans ce cas, des méthodes générales de paramétrisation, non spécifiques au problème traité, peuvent être préférées. C'est l'approche qui a été choisie dans [Gauvain et al., 1999] dans un problème de segmentation parole/musique d'émissions audiovisuelles. C'est également cette dernière approche qui a été finalement adoptée dans ce travail, après quelques efforts d'évaluation comparative d'une paramétrisation adaptée à la Reconnaissance Automatique du Locuteur [Magrin-Chagnolleau et al., 2000].

En Traitement du Signal Sonore, les paramètres acoustiques les plus utilisés sont empruntés au Traitement de la Parole, et sont essentiellement basés sur des méthodes de représentation temps-fréquence du signal. La bibliographie sur ces méthodes est assez vaste. Cependant, des survols peuvent être trouvés dans [Mammone et al., 1996, D'Alessandro et Demars, 1991, Davis et Mermelstein, 1980]. Pour un développement plus appronfondi des méthodes temps-fréquence, les ouvrages [Rabiner et Juang, 1993, Calliope, 1989, Lienard, 1977, Flanagan, 1972] constituent d'assez complètes références.

La méthodologie classique consiste à extraire régulièrement (par exemple chaque 10 ms) un ensemble (vecteur) de paramètres acoustiques, sur de très petites portions (trames) du signal, d'égale durée (par exemple 20). Le signal sonore est alors représenté par l'ensemble des vecteurs acoustiques ainsi extraits. Le processus d'extraction des vecteurs acoustiques est schématisé par la figure 1.2.

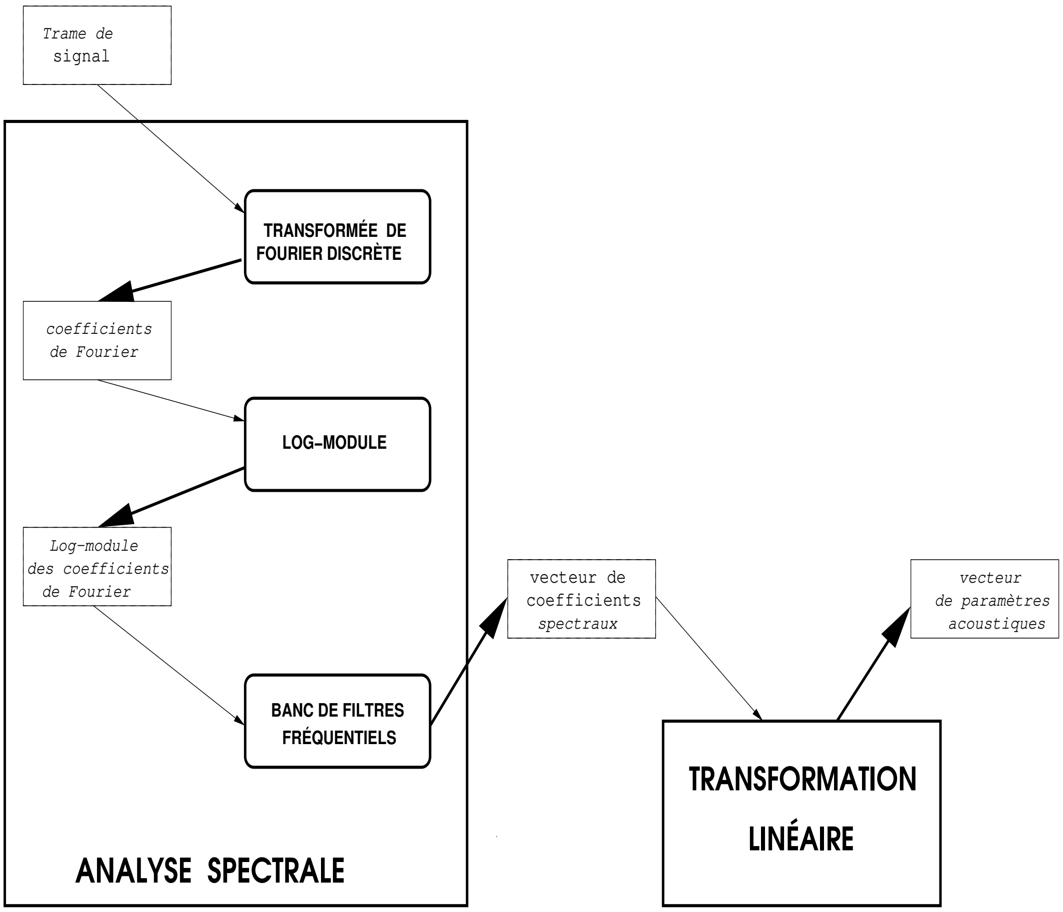

Fig. 1.2 – *Schéma du processus de paramétrisation acoustique*

Ce processus effectue d'abord une analyse du spectre de la trame, afin d'extraire un vecteur des paramètres spectraux (chaque paramètre correspondant à une fréquence). Le vecteur acoustique est ensuite obtenu par transformation affine du vecteur spectral. L'analyse spectrale et les transformations linéaires les plus utilisées vont être respectivement présentées dans les sections 1.1.1 et 1.1.2 suivantes.

1.1.1 Analyse spectrale

Le vecteur de coefficients spectraux de la trame est obtenu en mettant en cascade les trois opérations schématisées sur la figure 1.2.

1. Une Transformée de Fourier Discrète.
2. L'opérateur *log-module* est utilisé pour s'affranchir du spectre de phase considéré comme peu informatif en Traitement de la Parole. Et l'opérateur *log*, qui sert à se ramener à une échelle proportionnelle à celle des déciBels (dB).
3. Un lissage spectral en utilisant un banc de filtres disposés selon une échelle qui approxime l'échelle Mel. On utilise à cet effet, une échelle qui est linéaire en dessous de 1000 Hz et logarithmique au-delà, comme proposée dans [O'Shaughnessy, 1987]. Les filtres utilisés sont de type triangulaire comme ceux présentés dans [Davis et Mermelstein, 1980]. Ces derniers sont régulièrement espacés sur l'échelle Mel, et choisis de manière à couvrir toute la bande de fréquence du signal. L'utilisation de l'échelle Mel est optionnelle, bien que très fréquemment utilisée.

1.1.2 Transformations linéaires usuelles

Le vecteur de paramètres acoustiques est obtenu par transformation linéaire du vecteur de paramètres spectraux.

Les paramètres acoustiques les plus utilisés sont les coefficients cepstraux, qui sont obtenus par une transformée en cosinus. Ce consensus de la Communauté Scientifique Parole est d'abord due aux bonnes performances dans différentes applications comme la Reconnaissance Automatique de la Parole et la Reconnaissance Automatique du Locuteur, mais également, à quelques propriétés théoriques simples. Le cepstre est défini de manière plus générale, pour un signal à temps continu, comme étant la Transformée de Fourier Inverse du logarithme (complexe) de la Transformée de Fourier du signal. L'analyse cepstrale, introduite à l'origine par [Oppenheim, 1969], est basée sur une modélisation du signal comme étant la sortie d'un filtre linéaire, en réponse à une excitation. Alors dans le domaine cepstral (quéfrentiel), le cepstre du signal se décompose en la somme du cepstre du filtre et de celui de l'excitation. Cette relation additive rend plus facile la séparation entre la source et le filtre [Childers et al., 1969]. De plus, elle est à la base d'une technique de réduction de l'effet du canal de transmission en Traitement de la Parole : la *soustraction cepstrale* qui consiste à soustraire aux coefficients cepstraux leur moyenne à long terme sur tout le signal d'un enregistrement. Il est important de noter que cette technique, en plus de l'effet du canal (associé au filtre), soustrait également la moyenne cepstrale de l'excitation.

Deux autres transformations, issues de l'Analyse de Données, sont également utilisées : l'Analyse en Composantes Principales et l'Analyse Discriminante Linéaire. Ce sont des méthodes classiques d'Analyse de Données qui sont bien décrites par exemple dans [Saporta, 1990]. Les vecteurs acoustiques obtenus par ces méthodes peuvent être appelés les Composantes Principales du Spectre (CPS), utilisés par exemple dans [Magrin-Chagnolleau, 1997, Seck et al., 1998] et les Composantes Discriminantes du Spectre (CDS), expérimentées dans [Magrin-Chagnolleau et al., 2000]. Des extensions, qui consistent à effectuer ces transformations sur les vecteurs étendus (sur plusieurs trames consécutives) de coefficients spectraux, pour prendre en compte l'information dynamique au fil des trames, sont également utilisées [Magrin-Chagnolleau, 1997].

1.2 Modélisation

Dans cette section, le signal sonore est présenté sous forme paramétrisée. Il sera donc question d'un ensemble de n vecteurs de paramètres acoustiques $\{y_1, \cdots, y_n\}$. Ces vecteurs acoustiques sont supposés être extraits d'un ou de plusieurs signaux sonores, mais qui décrivent (pour guider l'intuition du lecteur) une classe de sons que l'on cherche à caractériser.

Comme annoncé dans l'introduction générale à ce travail, l'approche utilisée s'inspire considérablement de deux branches des Mathématiques Appliquées : les Probabilités et les Statistiques. La notion de distribution joue un rôle fondamental dans ces disciplines. Ainsi, un des éléments qui revient assez fréquemment dans ce travail est la distribution de probabilité d'un ensemble de vecteurs de paramètres acoustiques, qui a pour but de caractériser dans un sens probabiliste la classe de sons considérée.

Dans tout ce manuscrit, on suppose implicitement que les distributions admettent chacune une densité par rapport à la mesure de Lebesgue sur l'espace des vecteurs acoustiques (inclus dans $\mathbb{R}^d$, où d est la dimension des vecteurs acoustiques). On parlera alors de densité d'une distribution sans faire mention explicite de l'existence de cette densité.

En général la distribution exacte de l'ensemble des vecteurs acoustiques est inconnue. On est donc amené à estimer sa densité au moyen d'observations des vecteurs acoustiques $\{y_1, \cdots, y_n\}$, qui jouent ainsi le rôle d'exemples d'apprentissage. Le module de modélisation a pour fonction d'estimer la densité de la distribution d'un ensemble de vecteurs acoustiques. Ce module occupe une place de choix dans l'architecture des systèmes proposés dans ce travail.

On distingue généralement deux grandes familles de méthodes d'estimation de densités : les méthodes paramétriques et les méthodes non-paramétriques. On se place ici dans le cadre paramétrique, qui consiste à supposer que la densité que l'on cherche à estimer, appartient à une famille paramétrée de modèles, notée $\{p(.|\theta), \theta \in \Theta \subset \mathbb{R}^m\}$.

En Reconnaissance Automatique de la Parole, la famille de modèles qui fournit actuellement les meilleures performances, et qui est la plus utilisée, est celle des Modèles de Markov Cachés, dont les premières utilisations dans le domaine datent du milieu des années 1970 [Baker, 1975, Jelinek, 1976]. Ces modèles reposent sur une représentation de la distribution d'un signal sur un nombre fini d'états, dont chacun est défini par une distribution. La succession entre les différents états est alors prise en compte dans ces modèles, par la définition de probabilités de transition d'un état vers un autre. On parle de Modèle de Markov *Caché*, car il s'agit d'une situation où la variable d'état n'est pas observée (cachée).

Cette modélisation markovienne est adaptée à la reconnaissance d'un document sonore, structurée en une séquence de plusieurs classes de sons (classes acoustiques). C'est en effet le cas de la parole qui peut être vue comme une suite de mots, de phonèmes, etc... Par contre, dans d'autres tâches d'indexation sonore, l'enjeu majeur de la modélisation ne porte pas sur les transitions entre les différentes classes acoustiques. Par exemple, en Reconnaissance Automatique du Locuteur indépendamment du texte, le modèle du locuteur ne doit pas forcément rendre compte de l'ordre dans lequel les différents phonèmes sont prononcés, mais juste de la distribution globale des vecteurs acoustiques du locuteur. D'où l'idée des modèles de mélanges, qui conservent la décomposition sur un ensemble de distributions (états), mais font abstraction de l'ordre dans lequel les différents états sont visités. Le modèle de mélange peut également être vu (comme d'ailleurs toute distribution stationnaire) comme un Modèle de Markov à un seul état.

Les Modèles de Mélanges de Gaussiennes (MMG) constituent la famille de modèles utilisée dans ce travail.

Définition 1
Une variable aléatoire Y est un mélange de Gaussiennes, si et seulement si, il existe un entier K, des poids $c_1, \cdots, c_K$ appartenant à $]0,1]$ avec $\sum_{k=1}^{K} c_k = 1$, et K densités gaussiennes $g_1(.), \cdots, g_K(.)$ tels que la densité p de Y a pour expression :

$$p(y) = \sum_{k=1}^{K} c_k \; g_k(y) \tag{1.1}$$

Dans cette définition, K désigne le nombre de composantes du mélange, c_k est le poids associé à la $k^{\text{ème}}$ Gaussienne.

L'état de l'art actuel de la modélisation en Traitement du Signal Sonore, est basé sur cette famille de modèles. En plus des bonnes performances qu'elles fournissent, leur utilisation est également justifiée par l'existence d'outils de manipulation faciles à mettre en œuvre. Notamment la méthode d'estimation des MMG utilisée dans ce travail, l'algorithme Expectation-Maximization (EM) introduite par [Dempster et al., 1977]. Dans le cas de l'estimation Bayesienne des GMM, si la famille de distribution *a priori* est bien choisie, l'algorithme EM s'applique également assez aisément [Gauvain et Lee, 1994]. Les articles [Redner et Walker, 1984, Wu, 1983] sont de complètes références bibliographiques sur les propriétés et limites théoriques de l'alghorithme EM. L'estimation des paramètres d'un modèle de mélange en général est un problème qui a été l'objet de plusieurs travaux. Pour une revue bibliographique, le lecteur peut consulter par exemple les ouvrages [Titterington et al., 1985, Everitt et Hand, 1981].

Les MMG ont été introduits, en Vérification Automatique du Locuteur par [Reynolds, 1992], et depuis, constituent la famille de modèles de référence dans ce domaine. Pour un état de l'art, le lecteur peut consulter le journal [DSPJ, 2000], consacré aux évaluations NIST de 1999, qui portent sur le thème de la Reconnaissance Automatique du Locuteur en mode indépendant du texte. Ces modèles sont également utilisés dans d'autres tâches liées à l'indexation, comme pour la discrimination parole/musique dans [Gauvain et al., 1999, Scheirer et Slanley, 1997].

1.3 Calcul du score et décision

Le module de décision est la partie du système d'extraction de descripteurs, la plus dépendante de la tâche à accomplir. C'est la raison pour laquelle, il est relativement difficile d'en faire une description unifiée et indépendante de la tâche à accomplir par le système. Dans cette thèse, les problèmes sont en général formulés en termes de test statistique entre deux hypothèses. La statistique du test est alors le score du système et la décision est généralement prise par comparaison de ce score à un seuil. Le lecteur peut se référer à la partie correspondant au problème abordé, pour une formulation précise des hypothèses qui sont testées.

1.4 Conclusion

Ce chapitre a permis de présenter une achitecture modulaire des systèmes d'extraction de descripteurs étudiés et plus en détail la paramétrisation et la modélisation utilisée dans ce travail. Cette architecture donne une vue d'ensemble des différents systèmes en les réunissant sous la même structure générale. L'organisation modulaire permet de mieux identifier le rôle que peut jouer une partie du système, de l'évaluer individuellement, et éventuellement de la remplacer en vue d'améliorer le système. Dans la suite, la description et la spécification des systèmes vont s'appuyer sur l'architecture définie dans ce chapitre.

Chapitre 2

Test d'hypothèses

Les systèmes d'extraction de descripteurs étudiés dans ce travail sont généralement placés dans le cadre d'un test statistique entre deux hypothèses. L'objet de ce chapitre est de rappeler quelques notions sur les tests statistiques, en essayant de mettre l'accent sur le lien avec les systèmes. Après l'introduction de quelques notations et définitions dans la section 2.1, les sections 2.2 et 2.3 présentent les grandeurs caractéristiques usuelles pour chacun des deux types de tests que l'on distingue généralement : les tests d'hypothèses simples et les tests d'hypothèses composites.

2.1 Notations et définitions

On se propose de définir dans cette section l'essentiel des notions utilisées par les tests d'hypothèses statistiques. Soit $\mathcal{Y} = \{Y_i, i \geq 1\}$ un processus aléatoire à temps discret et à valeurs dans $\mathbb{R}^d$. On suppose que ce processus est une suite de variables aléatoires indépendantes dont la distribution de probabilité de chacune d'elle admet une densité[1] appartenant à la famille paramétrée $\{p(.|\theta), \quad \theta \in \Theta \subset \mathbb{R}^m\}$. Ici, $p(.|\theta)$ est une densité de distribution qui dépend d'un paramètre θ. Et Θ représente l'ensemble des paramètres de distribution, supposé être un sous-ensemble de $\mathbb{R}^m$. On note y_i une observation de la Y_i.

1. Par rapport à la mesure de Lebesgue.

Dans le contexte d'un système d'extraction de descripteurs d'un document sonore, $\{y_1,\cdots,y_n\}$ représente la séquence de vecteurs acoustiques extraits du document, et $\{Y_1,\cdots,Y_n\}$ est le modèle théorique de cette séquence.

Un test basé sur une séquence de n observations $\{y_1,\cdots,y_n\}$ consiste à décider entre deux hypothèses portant sur les distributions des v.a. $\{Y_1,\cdots,Y_n\}$. Généralement, on distingue deux types d'hypothèses : les hypothèses simples et les hypothèses composites respectivement explicitées par les définitions 2 et 3 ci-après.

Définition 2 (Hypothèse simple)
On appelle hypothèse simple une hypothèse du type $\{\theta^* = \theta_h\}$, où θ_h est un élément de l'ensemble des paramètres Θ.

Définition 3 (Hypothèse composite)
On appelle hypothèse composite une hypothèse du type $\{\theta^* \in \Theta_h\}$, où Θ_h est un sous-ensemble de l'espace des paramètres Θ, non réduit à un singleton.

On dispose maintenant de tous les éléments pour définir un test statistique entre deux hypothèses sur une séquence finie d'observations.

Définition 4 (Test)
Un test entre deux hypothèses, H_0 et H_1, sur la loi de paramètre θ^* génératrice d'une séquence d'observations $\{y_1,\cdots,y_n\}$, consiste à décider entre ces deux hypothèses.

Cette définition est schématisée par la figure 2.1. Sur ce schéma, on observe une réalisation $\{y_1,\cdots,y_n\}$ d'un processus aléatoire caractérisé par une distribution de paramètre θ^*. Deux hypothèses sont alors envisagées sur la distribution du processus : $H_0 = \{\theta^* \in \Theta_0\}$ et $H_1 = \{\theta^* \in \Theta_1\}$. Le test consiste à proposer une hypothèse estimée $\hat{H}$ parmi les deux hypothèses possibles H_0 et H_1.

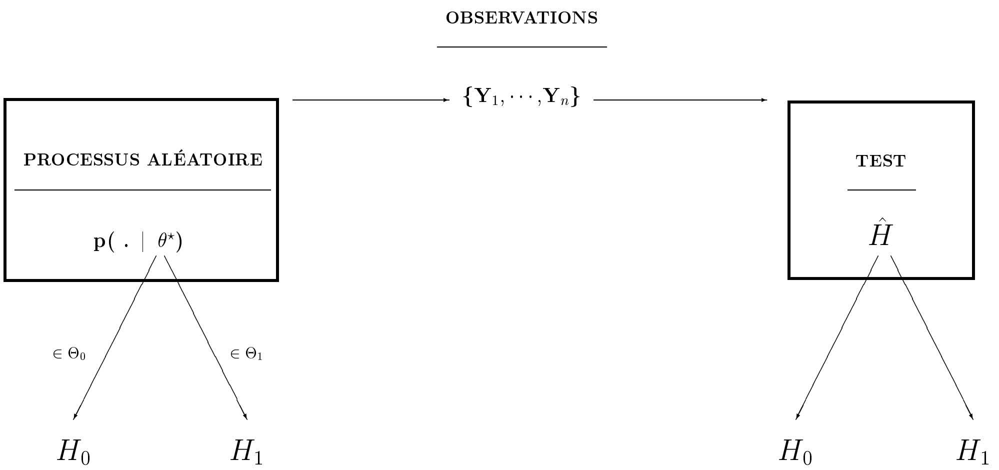

FIG. 2.1 – *Représentation schématique d'un test statistique.*

Usuellement, on choisit l'une des hypothèses, en général notée H_0, qu'on appelle *hypothèse nulle* et l'autre hypothèse, H_1, est appelée *hypothèse contraire.* Un test peut alors être défini par le sous-ensemble de $(\mathbb{R}^d)^n$, constitué par les séquences $\{y_1, \cdots, y_n\}$ pour lesquelles on rejette l'hypothèse l'hypothèse nulle H_0.

$$D_n = \Big\{ (y_1, \cdots, y_n) \in (\mathbb{R}^d)^n \ \text{ tel que le test rejette } H_0 \Big\} \tag{2.1}$$

La région de rejet D_n est aussi appellée *région critique du test.* Son complémentaire, noté D_n^c, est la région d'accpetation de H_0. On peut aussi définir un test par l'intermédiaire de la fonction indicatrice de sa région de rejet notée $\varphi_n = \varphi_n(y_1, \cdots, y_n)$, appelée *fonction de rejet du test*:

$$\begin{aligned} \varphi_n = \varphi_n(y_1, \cdots, y_n) &= \mathbf{1}_{\{(y_1, \cdots, y_n)\, \in\, D_n\}} \\ &= \begin{cases} 1 & \text{si} \quad (y_1, \cdots, y_n) \in D_n \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{2.2}$$

On se restreint dans ce travail aux tests dont la région de rejet est définie au moyen d'une statistique réelle $\Lambda_n = \Lambda_n(y_1, \cdots, y_n)$ par comparaison à un seuil λ. Une telle région de rejet est de la forme :

$$\{(y_1, \cdots, y_n) : \Lambda_n(y_1, \cdots, y_n) > \lambda\} = \{\Lambda_n(y_1, \cdots, y_n) > \lambda\} \tag{2.3}$$

Le rapport de vraisemblance entre un paramètre θ_1 de l'hypothèse H_1 et un paramètre θ_0 de l'hypothèse H_0 joue un rôle important dans la théorie statistique des tests. Il a pour expression :

$$\frac{p(y_1, \cdots, y_n|\theta_1)}{p(y_1, \cdots, y_n|\theta_0)} \tag{2.4}$$

Les exemples 1 et 2 énoncent des tests basés sur le rapport de vraisemblance. Le premier porte sur des hypothèses simples, et le second est une généralisation du premier aux hypothèses composites.

Exemple 1 (Test du rapport de vraisemblance)

Le test du rapport de vraisemblance (pour des hypothèses simples) de $H_0 = \{\theta^* = \theta_0\}$ contre $H_1 = \{\theta^* = \theta_1\}$, a pour statistique le rapport de vraisemblance, et est défini pour un seuil λ par sa région de rejet

$$\left\{\frac{p(y_1, \cdots, y_n|\theta_1)}{p(y_1, \cdots, y_n|\theta_0)} > \lambda\right\}$$

Exemple 2 (Test du rapport de vraisemblance généralisé)

Le test du rapport de vraisemblance, défini ci-desssus pour des hypothèses simples, se généralise aux cas d'hypothèses composites, $H_0 = \{\theta^* \in \Theta_0\}$ et $H_1 = \{\theta^* \in \Theta_1\}$, où Θ_0 et Θ_1 sont des sous-ensembles de Θ. Ce test est basé sur la statistique du rapport de vraisemblance généralisé qui a pour expression :

$$\frac{\sup_{\theta_1 \in \Theta_1} p(y_1, \cdots, y_n|\theta_1)}{\sup_{\theta_0 \in \Theta_0} p(y_1, \cdots, y_n|\theta_0)}$$

Le rapport de vraisemblance généralisé est le rapport entre le maximum de vraisemblance sous H_1 et celui sous H_0. C'est la statistique la plus utilisée dans le cas d'hypothèses composites. Le test basé sur cette statistique a pour région de rejet relatif à un seuil λ

$$\left\{ \frac{\sup\limits_{\theta_1 \in \Theta_1} p(y_1, \cdots, y_n|\theta_1)}{\sup\limits_{\theta_0 \in \Theta_0} p(y_1, \cdots, y_n|\theta_0)} > \lambda \right\}$$

Cette section a permis de définir quelques notions de bases très utilisées en théorie statistique des tests. Compte tenu du rôle important que joue les tests statistiques dans ce travail, ces notions reviendront assez souvent dans ce manuscrit.

2.2 Caractéristiques des tests d'hypothèses simples

Cette section rappelle les caractéristiques usuelles des tests entre deux hypothèses. Elles jouent également le rôle de critères de comparaison des tests. Deux types d'erreur sont possibles pour un test entre deux hypothèses : le faux rejet de H_0 ou sa fausse acceptation. Sur cette base, un test entre deux hypothèses simples est usuellement caractérisé par deux grandeurs : la probabilité de faux rejet et celle de fausse acceptation. Ces grandeurs sont définies de manière plus concise ci-après.

Définition 5 (Risque de première espèce)
On appelle risque de première espèce α d'un test de région de rejet D_n, sa probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie :

$$\alpha = \mathbb{P}_{\theta_0}[D_n]$$

Le risque de première espèce est aussi appelé niveau du test ou probabilité de fausse alarme.

Définition 6 (Risque de deuxième espèce)
On appelle risque de deuxième espèce β d'un test de région de rejet D_n, sa probabilité de rejeter l'hypothèse H_1 alors qu'elle est vraie :

$$\beta = \mathbb{P}_{\theta_1}[D_n^c]$$

Le risque de deuxième espèce est aussi appelé la probabilité de non-détection. La puissance du test est la probabilité d'accepter H_1 alors qu'elle est vraie, et correspond à la quantité $1-\beta$.

L'idéal pour un test est que ses deux types d'erreurs, le faux rejet et la fausse acceptation, soient les plus petits possibles. Pour un test d'hypothèses simples, il est souvent difficile de minimiser simultanément les deux types d'erreurs. On est alors amené à comparer deux tests sur la base de l'un des types d'erreur en fixant l'autre. Ainsi le critère usuel d'optimalité d'un test est la puissance $(1-\beta)$ à niveau fixé. Cette idée de comparaison des tests est formalisée par le critère d'optimalité défini ci-après.

Définition 7 (Plus Puissant (PP))
Un test de niveau α est dit Plus Puissant (PP) si, parmi tous les tests de niveau supérieur à α, il maximise la puissance.

Le théorème suivant énonce l'optimalité du test du rapport de vraisemblance, dans le cas d'hypothèses simples. Ce résultat est à la base du rôle que joue le rapport de vraisemblance dans les tests d'hypothèses.

Théorème 1 (Neyman-Pearson)
On considère le test du rapport de vraisemblance pour des hypothèses simples dont la région

de rejet est

$$\left\{\frac{p(y_1,\cdots,y_n|\theta_1)}{p(y_1,\cdots,y_n|\theta_0)} > \lambda\right\}$$

1. *Il existe un seuil λ, tel que le test soit de niveau α;*
2. *Si un test du r.v. est de niveau α pour un seuil λ, alors il est le PP parmi les tests de niveau supérieur à α;*
3. *Si un test est PP à un niveau α, alors il existe un seuil λ pour lequel le test du r.v. coincide p.s. avec le test PP; Ce test du r.v. est de niveau α, sauf s'il existe un autre test de niveau supérieur à α et de puissance 1.*

2.3 Caractéristiques des tests d'hypothèses composites

Les grandeurs caractéristiques des tests d'hypothèses composites sont très similaires à celles définies dans le cas d'hypothèses simples. Les sous-ensembles des paramètres correspondant à l'hypothèse nulle H_0 et l'hypothèse contraire H_1 seront respectivement notés Θ_0 et Θ_1. Toutes les grandeurs définies dans le cas d'hypothèses simples ont des analogues sous les mêmes appellations dans le cas d'hypothèses composites.

Définition 8 (Niveau)
Le niveau du test, noté α, est défini par :

$$\alpha = \sup_{\theta_0\in\Theta_0} \mathbb{P}_{\theta_0}[D_n]$$

Définition 9
La probabilité de non-détection du test $\beta(.)$ est définie en un point $\theta \in \Theta_1$ de l'hypothèse contraire H_1 par :

$$\beta(\theta_1) = \mathbb{P}_{\theta_1}[D_n^c]$$

Définition 10 (Uniformément Plus Puissant (UPP))
Un test de niveau α est dit Uniformément Plus Puissant (UPP), si parmi tous les tests de niveau supérieur à α, il maximise la puissance en tout point θ_1 de Θ_1.

2.4 Cadre asymptotique des tests

En conservant les mêmes notations que dans les précédentes sections, on s'intéresse au comportement asymptoptique de la suite de tests d'hypothèses composites, quand le nombre d'observations n varie. Pour n fixé, on note $\theta_n^\star$, le vrai paramètre de distribution des observations. Les hypothèses du test sont alors $H_0 = \{\theta_n^\star \in \Theta_0\}$ (nulle) et $H_1 = \{\theta_n^\star \in \Theta_1\}$ (contraire). Il existe deux cadres asymptotiques : l'Asymptotique Locale et l'Asymptotique Non-Locale, définis selon le comportement asymptotique du paramètre de distribution des observations $\theta_n^\star$, et sont précisés dans les sections qui suivent.

2.4.1 Asymptotique Locale

L'asymptotique locale considère une suite de paramètres θ_n^* telle que :

$$\sqrt{n}(\theta_n^* - \theta_0) \underset{n\to+\infty}{\longrightarrow} \delta$$

Il existe dans ce cas des suites de tests pour lesquelles les deux types d'erreurs admettent des limites non-nulles, qui peuvent alors être utilisées comme caractéristiques asymptotiques du test. On appelle niveau et erreur asymptotiques de la suite de test, les quantités respectives

$$\alpha = \sup_{\theta\in\Theta_0} \lim_{n\to\infty} \mathbb{P}_\theta[D_n] \tag{2.5}$$

$$\beta(\theta) = \lim_{n\to\infty} \mathbb{P}_\theta[D_n^c] \qquad \theta \in \Theta_1. \tag{2.6}$$

La quantité $1 - \beta(\theta)$ est appelée la puissance asymptotique.

Définition 11 (LAUPP)
Une suite de tests est dite Localement Asymptotiquement Uniformément Plus Puissant (LAUPP) au niveau α, si parmi toutes les suites de niveau asymptotique supérieur à α, elle maximise la puissance asymptotique en tout point de Θ_1.

Pour la détection de la présence d'un locuteur dans un document, un des systèmes proposés dans ce travail se ramène à tester $\{\theta_n^\star = 0\}$ contre $\{\theta_n^\star > 0\}$, où θ est la proportion (en durée) de présence du locuteur cible dans le document. Le théorème suivant donne sous certaines hypothèses de régularité, un test Localement Asymptotiquement Plus Puissant, pour tester $\{\theta_n^\star = \theta_0\}$ contre $\{\theta_n^\star > \theta_0\}$, pour toute valeur de θ_0

Théorème 2

Sous les conditions de régularité énoncées à l'annexe B.2, pour tester $\{\theta_n^\star = \theta_0\}$ contre $\{\theta_n^\star > \theta_0\}$, la suite de tests de régions de rejet

$$D_n \quad = \quad \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} \log p(y_i|\theta) \bigg|_{\theta=\theta_0} > \lambda_n \right\} \quad ,$$

où λ_n est choisi de manière que

$$\mathbb{P}_{\theta_0}[D_n] \quad = \quad \alpha \quad ,$$

est Localement Asymptotiquement Plus Puissant.

Une preuve de ce théorème peut être trouvée dans [Genot-Catalot et Picard, 1993] section 3.2. Remarquons au passage, que cette suite de tests est également Asymptotiquement Plus Puisssant, car pour tout $\theta_1 \in \Theta_1$, sa puissance asymptotique est maximale (égale à 1). En d'autres termes, sous la loi P_{θ_1},

$$\mathbb{P}_{\theta_1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p(Y_i|\theta) \bigg|_{\theta=\theta_0} > \lambda_n \right] \longrightarrow 1$$

2.4.2 Asymptotique Non-Locale

Pour l'asymptotique non-locale, la suite $\theta_n^\star$ satisfait :

$$\sqrt{n}(\theta_n^* - \theta_0) \underset{n \to +\infty}{\longrightarrow} +\infty$$

Dans ce cas, on peut trouver dans la plupart des cas des tests dont les deux types d'erreurs tendent vers 0. De telles suites de tests sont dites consistantes. Dans cette situation, on caractérise le test par les vitesses de convergence exponentielles des deux types d'erreurs. On définit le log-niveau et la log-erreur respectivement par :

$$\alpha = - \sup_{\theta_0 \in \Theta_0} \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{\theta_0}[D_n] \tag{2.7}$$

$$\beta(\theta_1) = - \lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}_{\theta_1}[D_n^c] \qquad \theta_1 \in \Theta_1. \tag{2.8}$$

Définition 12 (LAO)

Une suite de tests est dite Logarithmiquement Asymptotiquement Optimale (LAO) au log-niveau α, en $\theta_1 \in \Theta_1$, si parmi toutes les suites de log-niveau supérieur à α, elle maximise la log-erreur en tout point θ_1 de Θ_1.

Il est relativement difficile de trouver des tests LAO, car les caractéristiques font appel à des probabilités d'événements rares, qu'il n'est pas facile de manipuler, bien qu'il existe des outils théoriques dans le cadre thématique des Grandes Déviations sur lequel les ouvrages spécialisés [Shwartz et Weiss, 1995, Dembo et Zeitouni, 1993] constituent de bonnes références.

2.5 Conclusion

Dans ce chapitre, nous reprenons des concepts et résultats de base de la théorie statistique des tests entre deux hypothèses. Dans le cas d'hypothèses simples, il est connu que le théorème de Nerman-Pearson garantit l'optimalité du test du rapport de vraisemblance usuel. Pour des d'hypothèses composites, il est relativement difficile de trouver des tests optimaux quand le nombre d'observations n est fixé. Cependant, dans un cadre asymptotique quand $n \longrightarrow +\infty$, assez souvent, on peut trouver des tests pour lesquels les deux types d'erreurs tendent vers 0. Maintenant, si on compare de tels tests sur leur vitesse de convergence exponentielle, il n'est pas toujours aisé de trouver un test optimal facile à mettre en œuvre.

Deuxième partie

DÉTECTION DE CHANGEMENT DE PLAGE SONORE

Introduction

La détection et estimation de changement de comportement dans un signal consistent à localiser les instants où certaines caractéristiques du signal changent. Les dites caractéristiques du signal peuvent être aussi bien d'ordre numérique que qualitatif. Il peut s'agir par exemple, de détecter les moments de passage d'un locuteur à un autre dans une conversation, ou de localiser les transitions de la parole à la musique dans un document audiovisuel. De manière générale, les changements peuvent être transitoires ou instantanés. Dans ce travail, on s'intéresse spécifiquement aux changements instantanés ou supposés tels.

Nous plaçons ce problème dans un cadre statistique, où les changements instantanés sont alors appelés *ruptures*. La détection de rupture est au carrefour entre plusieurs thématiques de la statistique inférencielle : les tests d'hypothèses, la théorie de l'estimation, le contrôle de qualité, etc... Le contrôle de la qualité de produits manufacturés est une des applications à l'origine de la détection de ruptures [Shewart, 1931]. Les années 50 constituent une étape importante dans l'évolution de la formalisation du problème. On peut en l'occurence citer [Girshick et Rubin, 1952] et [Page, 1954], en relation avec les travaux de Shewart sur la détection automatique de défauts sur les produits manufacturés.

Depuis lors, la détection de ruptures a suscité un intérêt croissant auprès de la communauté en Mathématiques Appliquées et en Traitement du Signal, et la bibliographie sur le sujet est relativement vaste. Les ouvrages [Brodsky et Darkhovsky, 1993] (plutôt théorique) et [Basseville et Nikiforov, 1993] (plus ouvert aux aspects applicatifs) constituent des références incontournables sur les méthodes de détection et d'estimation de ruptures. On peut également mentionner l'existence de travaux spécialisés dans certains types de problèmes de détection de rupture, comme [Carlin et al., 1992, Stephens, 1994] sur les approches Bayesiennes, et [Deshayes et Picard, 1983, Csörgő et Horváth, 1997] portant sur les propriétés asymptotiques de statistiques de test d'existence d'une rupture.

On peut distinguer deux types de méthodes de détection de ruptures : les méthodes séquentielles (généralement orientées surveillance) et les méthodes non-séquentielles (plus spécifiquement dédiées à l'analyse *a posteriori*). La différence fondamentale entre ces deux familles réside dans la contrainte sur le délai de détection. Il doit être aussi court que possible pour les méthodes séquentielles, d'où une décision à chaque observation. Au contraire, les méthodes non-séquentielles (dans un but descriptif par exemple) ne sont pas forcément contraintes par l'urgence de la détection. Par conséquent, le traitement est effectué après la phase complète d'observations. Tandis que les méthodes non-séquentielles sont évaluées par leurs probabilités de non-détection et leur probabilité de fausse alarme, l'appréciation des performances des méthodes séquentielles doit également tenir compte du délai de détection.

Notons toutefois que la frontière entre ces deux types de méthodes n'est pas aussi claire, dans la mesure où des méthodes séquentielles peuvent être utilisées sur des problèmes non-séquentiels, et vice-versa. C'est le cas notamment de [Brandt, 1983] et de [Delyon et al., 1988, André-Obrecht, 1993], où des méthodes de détection séquentielle de ruptures sont appliquées à la segmentation de signaux de parole.

Le champ d'application couvert par la détection de ruptures s'est considérablement élargi. Les sujets rattachés à la surveillance sont de nature très diversifiés. En effet, celle-ci peut porter sur des infrastructures industrielles, des réseaux de télécommunications, de grandes structures (ponts, immeubles, sous-sol d'une région,...) pouvant être soumises à des vibrations ou aléas naturels, la médecine, etc... Les signaux biomédicaux (électroencéphalogrammes, électrocardiogrammes) ont fait l'objet de plusieurs applications de la détection de ruptures. Les applications aux signaux biomédicaux peuvent être orientées surveillance, quand il s'agit de veiller sur l'état physiologique d'un individu. Mais elles peuvent aussi être orientées analyse, lorsque le signal est segmenté en zones de stationnarité, dans un but descriptif.

D'autres applications existent dans des domaines en pleine expansion comme la parole, le son, l'image et le multimédia en général. La détection de rupture peut être utilisée comme méthode de segmentation temporelle, préalable à une phase de reconnaissance des segments. De récents travaux sur la segmentation d'une conversation en plages mono-locuteur (c'est-à-dire dans lesquelles seul un locuteur est présent) peuvent être trouvés dans [Beigi et Maes, 1998, Chen et Gopalakrishnan, 1998, Delacourt, 2000].

L'objet du premier chapitre de cette seconde partie est de donner des repères formels et méthodologiques sur la détection de ruptures dans un signal aléatoire. On introduit d'abord le formalisme nécessaire, puis on définit le type d'approche étudié. On termine par une présentation d'exemples de méthodes connues et qui sont des références dans la communauté Traitement du Signal.

Le chapitre 2 est consacré à la méthode de segmentation proposée dans ce travail. Il présente notamment l'utilisation d'une statistique de test de rupture qui peut s'interpréter comme une estimation empirique de la divergence de Kullback entre les distributions avant et après rupture. Une autre statistique de test similaire mais différente, également basée sur une estimation de la divergence de Kullback, a déjà été utilisée dans de précédents travaux [Basseville et Benveniste, 1983b], mais dans un cadre de test d'hypothèses séquentielles.

Le chapitre 3 est une étude comparative entre la statistique proposée dans ce travail et la statistique du rapport de vraisemblance généralisé, très utilisée pour la détection de rupture. La comparaison portera notamment sur leurs propriétés asymptotiques.

Chapitre 1

Détection et estimation de ruptures : formulation et techniques existantes

1.1 Formulation de la détection et estimation de ruptures

Soit $\mathcal{Y} = \{Y_1, \cdots, Y_n, \cdots\}$ un signal multidimensionnel à temps discret, à valeurs dans $\mathbb{R}^d$. Relativement aux tâches abordées dans ce travail, ce signal est celui des vecteurs de paramètres acoustiques, extraits des trames d'un signal sonore. On suppose qu'il existe des instants $r_1 < r_2 < \cdots$, tels que le signal est stationnaire sur chaque intervalle de temps $[r_i + 1, r_{i+1}]$, avec la convention $r_0 = 0$. L'homogénéïté dans le segment, hypothèse à laquelle il a été fait allusion dans l'introduction, est ainsi modélisée par la stationnarité du signal. Les instants r_i sont les instants de changement de distribution stationnaire du signal, et sont appelés instants de ruptures. On notera y_i une observation de la variable Y_i.

La détection et estimation de ruptures consiste à localiser les éventuels instants de ruptures sur une séquence d'observations $\{y_1, \cdots, y_n\}$ du signal $\mathcal{Y}$. Les critères de classification des problèmes de détection et estimation de ruptures sont multiples. Pour ne retenir que ceux qui nous paraîssent les plus essentiels, on peut citer :

- Le caractère séquentiel (au fur et à mesure qu'on observe le signal) ou non-séquentiel (après la phase d'observation du signal) du traitement.
- La connaissance *a priori* sur les distributions des segments :
  - distributions des segments appartenant à une famille finie connue,
  - distributions des segments appartenant à une famille paramétrée, décrite par un paramètre continu,
  - aucune information *a priori* sur les distributions des segments.
- La connaissance *a priori* sur la distribution des instants de ruptures :
  - instants de ruptures aléatoires,
  - instants déterministes,

Dans ce travail, le problème de la détection et de l'estimation des ruptures est placé dans le cadre où le traitement peut être séquentiel ou non, et les distributions des segments sont décrites par une famille paramétrée de densités $\{p(.|\theta); \theta \in \Theta \subset \mathbb{R}^m\}$. Dans nos expériences, la famille des Modèles de Mélanges de Gaussiennes est utilisée. Dans certains cas, l'ensemble des distributions des segments est réduit à un nombre fini de distributions. Par exemple, si les segments correspondent à des plages sonores d'un certain nombre de classes de sons (parole et musique par exemple), la distribution des paramètres acoustiques de chaque classe de sons peut être estimée à l'aide de données d'apprentissage.

La plupart des méthodes dont il sera question ne supposent pas de connaissance *a priori* sur la distribution des instants de ruptures. On mentionnera cependant un cas très utilisé en indexation de documents sonores : celui dans lequel la distribution du signal est modélisée par un Modèle de Markov Caché pour lequel un segment correspond à un séjour dans un état : cela induit ainsi une certaine distribution sur les instants de changements d'état, c'est-à-dire de rupture.

Les sections 1.2 et 1.3 présentent respectivement, et de manière succincte, quelques méthodes non-séquentielles et séquentielles de détection de rupture particulièrement connues en Traitement du Signal.

1.2 Méthodes non-séquentielles

Une méthode non-séquentielle consiste en une détection et estimation des éventuelles ruptures après une phase complète d'observation du signal. On peut distinguer deux types de méthodes : d'une part les approches globales qui tentent d'estimer conjointement et de manière directe les paramètres de rupture $(k,r_1,\cdots,r_k)$ à partir des obesrvations $\{y_1,\cdots,y_n\}$; et d'autre part les méthodes locales qui se proposent de se ramener localement au cas d'une seule éventuelle rupture et de traiter le signal au fur et à mesure. Les sections 1.2.1 et 1.2.2 présente des exemples de méthodes locales et globales.

1.2.1 Une approche non-séquentielle globale

Le signal $\{Y_t,\ t \geq 1\}$ et le signal indicateur des instants de ruptures[1] peuvent être respectivement compris comme la partie observable et la partie cachée d'un signal plus complet. Les approches globales utilisent un modèle de la variable complète et tentent d'estimer les instants de ruptures à partir des observations $\{y_1,\cdots,y_n\}$ du signal accessible.

Une approche globale et directe afin de résoudre le problème de la détection et estimation de ruptures consiste, dans un premier temps, à effectuer un test de l'hypothèse d'absence de rupture contre l'hypothèse de présence de rupture. Dans un second temps, si la décision du test est en faveur d'une présence de rupture, on entame une phase d'estimation des instants de rupture. Ci-après, les énoncés exacts des hypothèses de présence de rupture H_0, et d'absence de rupture H_1 :

1. Il s'agit du signal $\{X_t, t \geq 1\}$ défini par : $X_t = 1$ si t est un instant de rupture et $X_t = 0$ sinon.

- H_0 : *Il existe* $\theta_0^\star \in \Theta$ *tel que* $Y_1,\cdots,Y_n$ *suivent la même distribution de paramètre* $\theta_0^\star$.
- H_1 : *Il existe un entier* $k^\star \geq 1$, $k^\star$ *instants* $1 \leq r_1^\star < \cdots < r_{k^\star}^\star < n$, *et* $(k^\star + 1)$ *paramètres de distribution* $\{\theta_1^\star,\cdots,\theta_{k^\star+1}^\star\}$, $\theta_i^\star \neq \theta_{i+1}^\star$, *tels que sur chaque intervalle* $[r_{i-1}^\star + 1, r_i^\star]$, *les variables* $\{Y_{r_{i-1}^\star+1},\cdots,Y_{r_i^\star}\}$ *suivent la même distribution de paramètre* $\theta_i^\star$. *On adoptera la convention* $r_0^\star = 0$ *et* $r_{k^\star+1}^\star = n$.

L'étoile ($\star$) signifie qu'il s'agit des vrais paramètres de distribution, $k^\star$ représente le nombre de ruptures, $r_i^\star$ désigne un instant de rupture, et $\theta_i^\star$ correspond à un paramètre de distribution d'un segment.

Les vrais paramètres n'étant pas fixés, les hypothèses H_0 et H_1 sont de types composites, justifiant généralement l'utilisation d'un test du rapport de vraisemblance généralisé. Deux cas sont alors envisagés, selon que l'on dispose ou non d'une distribution *a priori* sur les instants de ruptures.

En l'absence de distribution *a priori* sur les instants de ruptures

Le test du rapport de vraisemblance généralisé de H_0 contre H_1 est basé sur la statistique :

$$\frac{\displaystyle\sup_{\substack{(k,r_1,\cdots,r_k)\,\in\,\Gamma \\ (\theta_1,\cdots,\theta_{k+1})\,\in\,\Theta^{k+1}}} p(y_1,\cdots,y_n|k,r_1,\cdots,r_k,\theta_1,\cdots,\theta_{k+1})}{\displaystyle\sup_{\theta\in\Theta}\; p(y_1,\cdots,y_n|\theta)} \tag{1.1}$$

où k représente le nombre de ruptures, r_i désigne le $i^{\text{ème}}$ instant de rupture, Γ est l'ensemble des paramètres de ruptures $(k,r_1,\cdots,r_k)$ décrivant H_1, et θ_i correspond au paramètre de distribution du $i^{\text{ème}}$ segment. Le dénominateur de la statistique du rapport de vraisemblance généralisé (formule 1.1) est le maximum de vraisemblance sous H_0 (sachant qu'il n'y a pas de rupture), et le numérateur est le maximum de vraisemblance sous l'hypothèse H_1 (sachant qu'il y a au moins une rupture). Dans la pratique les paramètres $(k,r_1,\cdots,r_k)$ qui réalisent le maximum du numérateur, sont pris comme estimateurs des paramètres de rupture (lorsque le test décide qu'il y a une rupture).

En général, on se place dans le cas d'une famille de distributions des segments pour laquelle les estimateurs du maximum de vraisemblance sont accessibles. Le maximum de vraisemblance sous H_0 peut alors être obtenu relativement facilement.

Le maximum de vraisemblance sous H_1 est plus difficile à calculer. D'abord, il est clair que si Γ est l'ensemble des paramètres de rupture possibles, plus k est grand, plus on dispose de paramètres et plus le maximum de vraisemblance a tendance à augmenter, et l'optimum est atteint pour le plus grand nombre de ruptures $k = n - 1$ et $r_i = i$. D'ou l'intérêt de restreindre le domaine des paramètres de ruptures à un ensemble Γ. Il peut être construit en introduisant une contrainte (directe) sur k, ou (indirectement) sur la longueur des segments (écarts $(r_{i+1} - r_i)$). Ces contraintes peuvent également se justifier par le souci d'avoir un nombre suffisant d'observations sur un segment, pour une bonne estimation du paramètre de distribution du segment.

En présence d'une distribution *a priori* sur les instants de ruptures

Les contraintes sur les paramètres de rupture $(k,r_1,\cdots,r_k)$ peuvent également s'exprimer par le biais d'une distribution *a priori*, plaçant ainsi le problème dans un cadre Bayesien. Si $p(k,r_1,\cdots,r_k|\theta_1,\cdots,\theta_{k+1})$ dénote la densité *a priori* de $(k,r_1,\cdots,r_k)$ sous H_1, le rapport de vraisemblance généralisé devient (au lieu de la formule 1.1) :

$$\frac{\displaystyle\sup_{\substack{(k,r_1,\cdots,r_k)\,\in\,\Gamma \\ (\theta_1,\cdots,\theta_{k+1})\,\in\,\Theta^{k+1}}} \left\{ \begin{array}{c} p(y_1,\cdots,y_n|k,r_1,\cdots,r_k,\theta_1,\cdots,\theta_{k+1}) \\ *\ p(k,r_1,\cdots,r_k|\theta_1,\cdots,\theta_{k+1}) \end{array} \right\}}{\displaystyle\sup_{\theta\in\Theta}\ p(y_1,\cdots,y_n|\theta)} \tag{1.2}$$

Des travaux sur les approches Bayesiennes peuvent être trouvés par exemple dans [Carlin et al., 1992, Stephens, 1994].

En plus des difficultés mentionnées dans le cas d'absence de distribution sur les paramètres de rupture, la maximisation de la vraisemblance sous H_1 est un problème d'optimisation avec des paramètres discrets $(k,r_1,\cdots,r_k)$, sur un ensemble dont la taille peut croître très rapidement en fonction du nombre d'observations n et selon la structure de Γ. Des solutions existent. On peut par exemple utiliser des algorithmes stochastiques comme le recuit simulé, décrit dans [Anily et Federgruen, 1968, van Laarhoven et Aarts, 1987, Azencott, 1988]. Mais ces techniques d'optimisation sont très gourmandes en temps de calcul. Cela compromet l'application et l'évaluation à grande échelle de ces méthodes.

Il existe un cas de figure très utilisé en indexation de documents sonores, pour lequel la maximisation de la vraisemblance sous H_1 peut être obtenue plus aisément. Il s'agit du cas où le signal $\{Y_t,\ t \geq 1\}$ est modélisé par une Chaîne de Markov Cachée (ou partiellement observée). Un segment correspond au séjour dans un des états, et les distributions des segments sont décrites par une famille finie de K densités $\{p(.|\theta), \theta \in \Theta = \{\underline{\theta}_1,\cdots,\underline{\theta}_K\} \subset \mathbb{R}^m\}$.

La modélisation par un modèle de Markov Caché commence à être utilisée en indexation de signaux sonores, quand le signal est constitué de plages sonores appartenant à des classes de sons connues, le segment représentant un intervalle de séjour dans une des classes sonores. À ce titre, on peut faire référence aux récents travaux de [Gauvain et al., 1999] pour la segmentation de journaux télévisés en plages de parole, de musique et de silence, et ceux de [Sönmez et al., 1999, Meignier et al., 2000] pour la segmentation d'une conversation en plages appartenant à un seul locuteur.

Dans le cas d'un Modèle de Markov Caché, la densité des paramètres de rupture $(k,r_1,\cdots,r_{k+1})$ connaissant la succession d'états $(\theta_1,\cdots,\theta_{k+1})$, noté $p(k,r_1,\cdots,r_{k+1}|\theta_1,\cdots,\theta_{k+1})$, est donnée par :

$$p(k,r_1,\cdots,r_k|\theta_1,\cdots,\theta_{k+1}) \quad = \quad p_1\ (p_{1,1})^{r_1-1} \quad * \quad \prod_{j=2}^{k+1} p_{j,j-1}\ (p_{j,j})^{r_j-r_{j-1}-1} \tag{1.3}$$

où p_i est la probabilité *a priori* de l'état associé à la distribution de paramètre θ_i, et $p_{i,j}$ est la probabilité de transition de la distribution de paramètre θ_i à celle de paramètre θ_j. Dans ce cas, les paramètres de rupture qui maximisent la vraisemblance sous H_1, peuvent être obtenus au moyen de l'algorithme de Viterbi [Viterbi, 1967].

1.2.2 Une approche non-séquentielle locale

Pour remédier aux difficultés de mise en œuvre des approches globales, surtout quand la famille de distributions des segments est complexe, une approche locale consiste à se ramener à un traitement par blocs d'observations, sur lesquels le nombre maximum de ruptures présumé est raisonnable. Généralement, l'on se ramène à tester l'existence d'une rupture à chaque instant. La détection en un instant t donné peut être basée, aussi bien sur des observations au voisinage de t, que sur toute la séquence d'observations $\{y_1, \cdots ,y_n\}$.

La plupart des méthodes commencent par calculer, à chaque instant t, un *indice* (statistique) de rupture, puis à en extraire un *critère* de décision. La décision de présence d'une rupture en un instant est prise par comparaison du critère à un seuil. Ce principe de calcul en deux temps (indice puis critère) est à la base de nombreuses méthodes de segmentation de signaux. Des travaux récents en segmentation de documents sonores peuvent être trouvés dans [Montacié et Caraty, 1997, Beigi et Maes, 1998, Delacourt et al., 1999]. Des résultats théoriques sur cette approche sont présentés dans [Bertrand, 2000], pour des ruptures correspondant à un changement de moyenne dans une famille Gaussienne de distributions des segments.

La propriété fondamentale requise pour un indice de rupture est qu'elle prenne ses plus grandes valeurs aux instants de rupture. On utilise alors généralement comme indice, une statistique de test d'existence d'une rupture à chaque instant, sur une fenêtre de contexte autour de l'instant considéré, censée ne contenir au plus qu'une seule rupture : par exemple, une fenêtre de contexte, supposée de taille égale de part et d'autre de l'instant t, c'est-à-dire w observations $\{y_{t-w+1}, \cdots ,y_t\}$ pour le *passé* et w autres observations $\{y_{t+1}, \cdots ,y_{t+w}\}$ pour le *futur*, comme le schématise la figure 1.1.

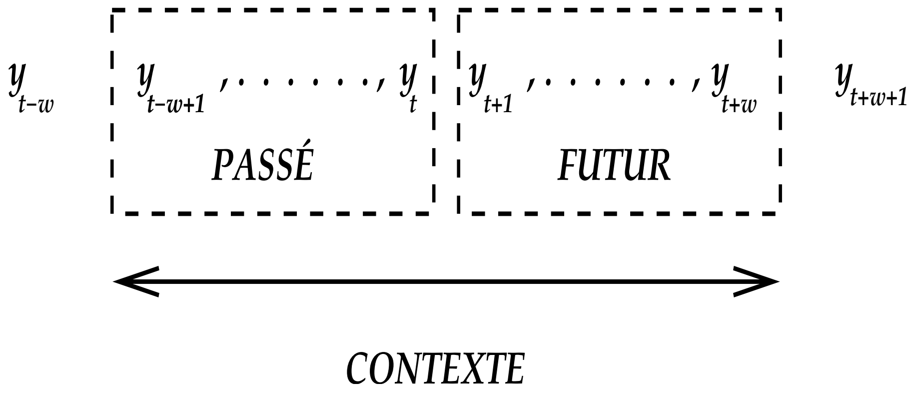

FIG. 1.1 – *Positions des fenêtres de contexte passé et futur d'un instant t pour le calcul d'une statistique de détection de rupture.*

Une des statistiques les plus utilisées est le rapport de vraisemblance généralisé du test d'existence de rupture en un instant t fixé. Cette statistique a pour expression en t :

$$\frac{\sup_{\theta_- \in \Theta} p(y_{t-w+1}, \cdots, y_t | \theta_-) \sup_{\theta_+ \in \Theta} p(y_{t+1}, \cdots, y_{t+w} | \theta_+)}{\sup_{\theta_0 \in \Theta} p(y_{t-w+1}, \cdots, y_{t+w} | \theta_0)}$$

La figure 1.2 illustre un exemple de rapport de vraisemblance généralisé (exemple repris plus en détail dans le chapitre suivant). Les maxima locaux de la statistique du r.v.g sont au voisinage des vrais instants de ruptures, qui se produisent toutes les 5 secondes. Cependant, cet exemple met clairement en évidence le fait que l'extraction de ces maxima locaux ne peut pas être obtenu par un seuillage direct de l'indice.

Plus généralement, la statistique de détection de rupture présente une allure très irrégulière, avec de nombreux maxima locaux voisins les uns des autres. C'est pourquoi, prendre une décision à chaque instant, en comparant la valeur de la statistique à un seuil, expose à un important risque de fausses alarmes. D'où le besoin d'extraire de cette statistique un critère de rupture moins sujet aux fausses alarmes, quand la décision est prise par

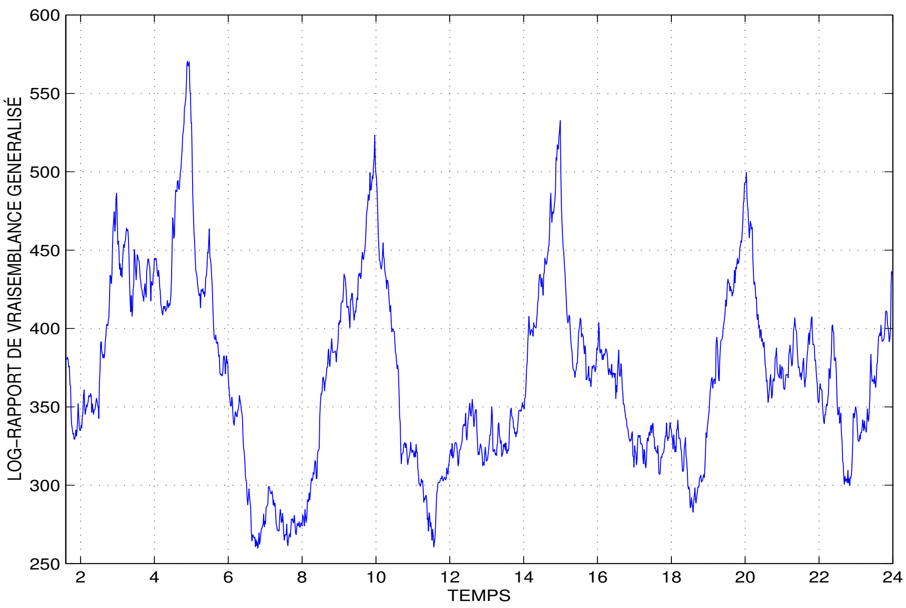

FIG. 1.2 – *Logarithme du rapport de vraisemblance généralisé instantané, sur une fenêtre de contexte glissante autour de l'instant courant. Les ruptures sont générées en concaténant des morceaux de parole et de musique de* 5 *secondes chacun. Les fenêtres du passé et du futur ont une durée de* 1,6 *s. Les vecteurs acoustiques utilisés sont constitués des* 25 *premières Composantes Principales du Spectre (CPS), et les modèles du passé et du futur sont Gaussiens.*

comparaison à un seuil.

Plusieurs méthodes d'extraction de critères peuvent être trouvées dans la littérature. Cependant, elles sont souvent empiriques et dépendantes de seuils et/ou de lissages intermédiaires plus ou moins explicitement formulés.
Le critère de décision proposé au chapitre 2 permet d'éviter les seuils et les lissages intermédiaires. L'approche qui y est décrite pour l'extraction du critère de décision, est basée sur la notion de hauteur d'un maximum, empruntée à l'optimisation stochastique. Cette méthode est évaluée sur une tâche de détection de rupture sur un signal obtenu par concaténation de réels segments de parole et de musique. Les expériences de l'évaluation en question sont présentées à la section 2.3.

1.3 Méthodes séquentielles

Une méthode séquentielle de détection de rupture est une méthode qui, à chaque instant d'observation, rend une décision d'absence ou de présence d'une rupture (avec éventuellement un certain retard, ou délai intrinsèque). Les méthodes séquentielles sont plus adaptées à certains problèmes de surveillance, bien qu'elles s'appliquent aussi à d'autres problèmes de détection de rupture. Le cadre théorique de ces méthodes est celui des tests séquentiels, présenté par exemple dans [Siegmund, 1985, Basseville et Nikiforov, 1993].

La séquentialité du traitement n'est pas une préoccupation de ce travail. Nous allons mentionner le principe de deux algorithmes séquentiels de détection de rupture, très connus dans la littérature : l'algorithme de Brandt [Brandt, 1983] et l'algorithme de la divergence [André-Obrecht, 1993, Basseville et Nikiforov, 1993, Delyon et al., 1988].

Soit $\hat{r}_0$ le dernier instant de rupture estimé. Ces algorithmes testent à partir de l'instant $\hat{r}_0$, l'existence d'une rupture à chaque nouvel instant d'observation. Si le test décide à l'instant t qu'il y a eu une rupture, on entame une procédure d'estimation d'une rupture entre $\hat{r}_0$ et t. Soit $\hat{r}_1$ le nouvel instant de rupture estimé. On recommence la même procédure de détection d'une rupture à partir de $\hat{r}_1$, et ainsi de suite.

1.4 Conclusion

Ce chapitre avait pour objet premier de donner un aperçu général des problèmes de détection et estimation de ruptures sur un signal aléatoire à temps discret. Et ce, afin de situer le cadre formel couvert par cette étude. Généralement on distingue deux types de méthodes : les algorithmes séquentiels et ceux non-séquentiels. Les applications visées par ce travail ne sont pas contraintes à une séquentialité du traitement. Ainsi, l'accent a été mis sur les méthodes non-séquentielles.

En indexation de documents audiovisuels, la détection et estimation de ruptures est souvent utilisée dans une tâche visant à reconnaître les segments. La plupart des travaux existants sur la segmentation évaluent la proportion (en durée) des plages correctement reconnues, au lieu du taux de frontières de segments correctement détectées et du taux d'omission.

Chapitre 2

Méthode proposée

L'objet de ce chapitre est de détailler la méthode de détection de ruptures statistiques sur un signal multidimensionnel, proposée dans ce travail. Le formalisme et les notations sont identiques à ceux introduits au début de la présente partie (section 1.1). La première section de ce chapitre présente les motivations de l'approche proposée, dont la description est effectuée à la section 2.2. La section 2.2.1 aborde le cas particulier d'une famille finie de distributions des segments, et la (dernière) section 2.3 présente des expériences en détection de ruptures de type parole/musique.

2.1 Motivations

Les méthodes de détection de ruptures statistiques abordées dans le précédent chapitre reposent en général sur une statistique de détection de rupture calculée à chaque instant, suivie de l'extraction d'un critère à partir duquel sont estimés les instants de rupture. La statistique de détection de rupture doit posséder de bonnes propriétés, lesquelles sont le plus souvent garanties grâce à la comparaison des observations avant et après l'instant courant. Le calcul du critère de rupture permettant l'extraction des instants de rupture est très souvent la phase la plus délicate à mettre en œuvre.

Un inconvénient des deux algorithmes séquentiels de détection de ruptures, celui de

Brandt et celui de la divergence (mentionnés à la section 1.3), est la dépendance causale des estimations des instants de rupture. En d'autres termes, si $\{\hat{r}_i, i \geq 1\}$ est la suite des instants de rupture estimés, $\hat{r}_i$ dépend de $\hat{r}_{i-1}$. Cette causalité dans le processus de détection présente le risque de propager les éventuelles erreurs de détection. D'où l'intérêt de rendre les estimations des instants de rupture indépendantes les unes des autres.

Certaines méthodes utilisent des paramètres (de réglage) intermédiaires qui interviennent de manière complexe dans le procédé de détection. C'est le cas par exemple, du biais dans l'algorithme de la divergence, du facteur de pénalité du BIC (Bayesian Information Criterion) utilisé dans [Chen et Gopalakrishnan, 1998, Delacourt et al., 1999], de certains coefficients de lissage de l'indice, de seuils de sélection, etc.... Il devient alors difficile d'effectuer le réglage (souvent nécessaire) du paramètre en question, dans le but de l'adapter au type de signaux et de ruptures traités.

Afin de s'affranchir de cet inconvénient, une solution consiste à se ramener à un critère n'utilisant aucun paramètre intermédiaire, que l'on comparera ensuite à chaque instant à un seuil pour prendre la décision finale de présence ou d'absence de rupture. La méthode de détection de rupture décrite à la section 2.2 a pour principales motivations de palier les inconvénients qui viennent d'être soulignés.

2.2 Description de la méthode de base proposée

La méthode de détection de rupture proposée consiste d'abord à calculer à chaque instant t, une statistique indicatrice de rupture, appelée indice de rupture. On note par $I(t)$ la valeur de cet indice à l'instant t, pour $t = 1, \cdots, n$. À partir de l'indice I, un critère de présence de rupture, également défini à chaque instant, est calculé. $C(t)$ désigne la valeur du critère à l'instant t. La décision de présence d'une rupture est prise à chaque instant t par comparaison de la valeur du critère $C(t)$ à un seuil λ.

La propriété fondamentale requise pour un indice de rupture est qu'elle prenne ses plus grandes valeurs aux (voisinages des) instants de rupture, et que ses valeurs aux instants éloignés des instants de rupture soient les plus faibles possibles. Relativement à cette propriété, le critère a pour fonction de bien mettre en evidence les maxima dominants de l'indice, en anéantissant les pics secondaires.

La section 2.2 présente les deux indices de rupture étudiés et la section 2.2 décrit la méthode d'extraction du critère de rupture utilisée.

Indices de rupture étudiés

Les deux indices de rupture étudiés correspondent au calcul, en un instant donné t, d'une dissimilarité (au sens large du terme) entre les observations dites du *passé* $\{y_{t-w+1}, \cdots, y_t\}$, et celles dites du *futur* $\{y_{t+1}, \cdots, y_{t+w}\}$, où w représente la taille (commune) du passé et du futur. Ces deux indices sont :

$$I_1(t) = \frac{1}{w} \log \frac{p(y_{t-w+1}, \cdots, y_t|\hat{\theta}_{t_-})\, p(y_{t+1}, \cdots, y_{t+w}|\hat{\theta}_{t_+})}{p(y_{t-w+1}, \cdots, y_{t+w}|\hat{\theta}_t)} \tag{2.1}$$

$$I_2(t) = \frac{1}{w} \left[\log \frac{p(y_{t-w+1}, \cdots, y_t|\hat{\theta}_{t_-})}{p(y_{t-w+1}, \cdots, y_t|\hat{\theta}_{t_+})} + \log \frac{p(y_{t+1}, \cdots, y_{t+w}|\hat{\theta}_{t_+})}{p(y_{t+1}, \cdots, y_{t+w}|\hat{\theta}_{t_-})} \right] \tag{2.2}$$

où :

$$\begin{cases} \hat{\theta}_{t_-} = \arg\max\limits_{\theta\in\Theta} p(y_{t-w+1}, \cdots, y_t|\theta) \quad \hat{\theta}_{t_+} = \arg\max\limits_{\theta\in\Theta} p(y_{t+1}, \cdots, y_{t+w}|\theta) \\ \hat{\theta}_t = \arg\max\limits_{\theta\in\Theta} p(y_{t-w+1}, \cdots, y_{t+w}|\theta) \end{cases}$$

La figure 2.1 représente en un instant t donné, les positions des fenêtres d'estimation des modèles du passé $\hat{\theta}_{t_-}$, du futur $\hat{\theta}_{t_+}$ et du contexte $\hat{\theta}_t$.

Cette approche par comparaison entre le passé et le futur au voisinage de l'instant courant t est couramment utilisée en détection de rupture statistique sur un signal. L'indice $I_1(t)$ correspond au logarithme (moyen) de la statistique du rapport de vraisemblance

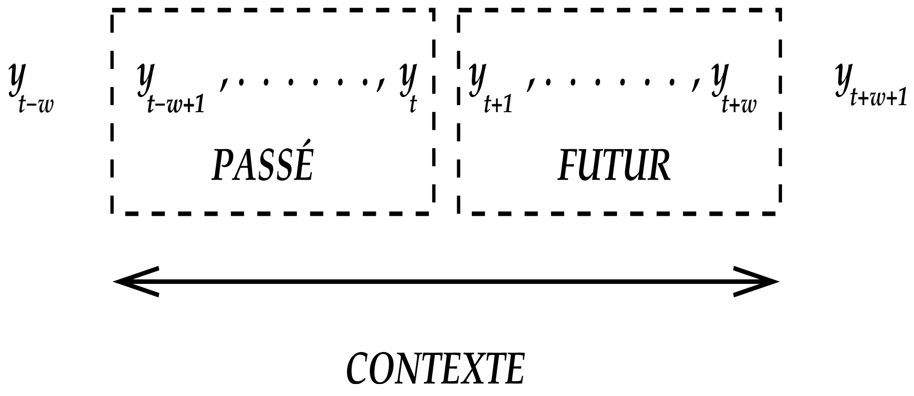

FIG. 2.1 – *Positions, à un instant t, des fenêtres d'estimation des modèles du passé $\hat{\theta}_{t_-}$, du futur $\hat{\theta}_{t_+}$ et du contexte $\hat{\theta}_t$, pour le calcul des indices de détection de rupture.*

généralisé, pour le test d'une rupture à l'instant t, basé sur les observations du contexte $\{y_{t-w+1},\cdots,y_{t+w}\}$. Quant à $I_2(t)$, c'est une approximation de la divergence de Kullback[1] entre le modèle ajusté au passé $\{y_{t-w+1},\cdots,y_t\}$ et celui ajusté au futur $\{y_{t+1},\cdots,y_{t+w}\}$. L'indice I_2 présente l'avantage de ne reposer que sur deux modèles ($\hat{\theta}_{t_-}$ et $\hat{\theta}_{t_+}$), au lieu des trois modèles ($\hat{\theta}_t$, $\hat{\theta}_{t_-}$ et $\hat{\theta}_{t_+}$) utilisés par I_1. Le chapitre 3 présente quelques propriétés statistiques des indices I_1 et I_2.

Pour le bon déroulement de l'algorithme, la taille w des fenêtres de contexte doit être suffisamment grande afin qu'il y ait assez d'observations pour bien estimer les modèles. Tout de même, w ne doit pas être trop grande, pour éviter d'avoir des observations hétérogènes en distribution pour l'estimation d'un modèle. Dans la pratique, une fenêtre trop grande est moins génante qu'une trop petite fenêtre, sauf dans le cas critique où le passé et le futur correspondent aux mêmes distributions hétérogènes.

1. La divergence de Kullback entre la distribution de paramètre θ_1 et celle de paramètre θ_2 est égale à : “$\mathbb{E}_{\theta_1}\left[\log\frac{p(Y|\theta_1)}{p(Y|\theta_2)}\right]+\mathbb{E}_{\theta_2}\left[\log\frac{p(Y|\theta_2)}{p(Y|\theta_1)}\right]$”. C'est une quantité positive qui mesure la dissemblance (dans un sens probabiliste) entre les deux distributions, qui n'est nulle que si elles sont les mêmes.

Extraction du critère de rupture

Une propriété idéale d'un indice de rupture est que ses maxima correspondent exactement aux instants de ruptures. Mais, dans la plupart des cas, il est difficile de trouver un indice avec une telle propriété. En général, l'indice de rupture dont on dispose a un comportement très irrégulier, comme le montre l'exemple de la figure 1.2 (section 1.2.2). Ainsi, la décision de présence ou non d'une rupture ne peut pas être prise à un instant t en comparant directement la valeur de l'indice $I(t)$ à un seuil, à cause du risque de fausses alarmes, notamment aux voisinages des maxima de l'indice. Nous utilisons l'approche suivante, décrite ci-après, en s'appuyant sur la figure 2.2.

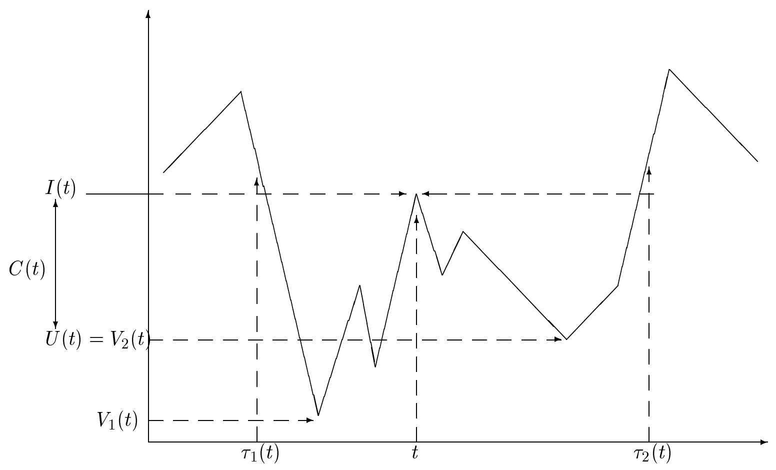

FIG. 2.2 – *Schéma d'extraction du critère de décision à partir d'un indice de rupture.*

Pour un instant t donné, on recherche les premiers instants $\tau_1(t) < t$ et $\tau_2(t) > t$, tels que :

$$I(\tau_1(t)) > I(t) \ \ et \ \ I(\tau_2(t)) > I(t) \tag{2.3}$$

Puis on calcule les quantités $V_1(t)$ et $V_2(t)$:

$$V_1(t) = \min_{\tau_1(t)<i<t} I(i) \quad \text{et} \quad V_2(t) = \min_{t<i<\tau_2(t)} I(i) \tag{2.4}$$

Et enfin :

$$U(t) = \max\{V_1(t),V_2(t)\}$$

Le critère de présence de rupture à l'instant t est alors défini par :

$$C(t) = I(t) - U(t)$$

Le critère C ainsi défini est positif aux instants qui sont maxima locaux de I, et est nul aux autres instants.

$U(t)$ est la plus haute des deux vallées les plus basses de part et d'autre de la région de la courbe dominée par $I(t)$. En ce sens, le critère de rupture $C(t)$ est une mesure de la dénivelation de l'indice $I(t)$, par rapport à un niveau de palier local.

La figure 2.3 représente un exemple d'indice de rupture (en haut) et le critère (en bas) qui a été extrait de l'indice. Il s'agit de l'indice du log-rapport de vraisemblance généralisé, calculé sur le même signal que l'exemple présenté sur la figure 1.2, à la section 1.2.2.

A partir du critère ainsi obtenu, la décision de présence où non d'une rupture à l'instant t, peut être prise en comparant directement la valeur du critère à un seuil :

$$C(t) \quad \underset{NONRUPT}{\overset{RUPT}{\gtrless}} \quad \lambda \tag{2.5}$$

Un traitement exactement dual peut être appliqué aux minima locaux lorsque l'on veut obtenir un critère qui met en évidence les extrema locaux en général au lieu des seuls maxima locaux. Ceci est détaillé un peu plus loin à la section 2.2.1 - *Application de la méthode de base.*

2.2.1 Cas d'une famille finie de distributions des segments

La méthode de détection de ruptures présentée à la section 2.2 sera appelée *méthode de base* dans la suite.

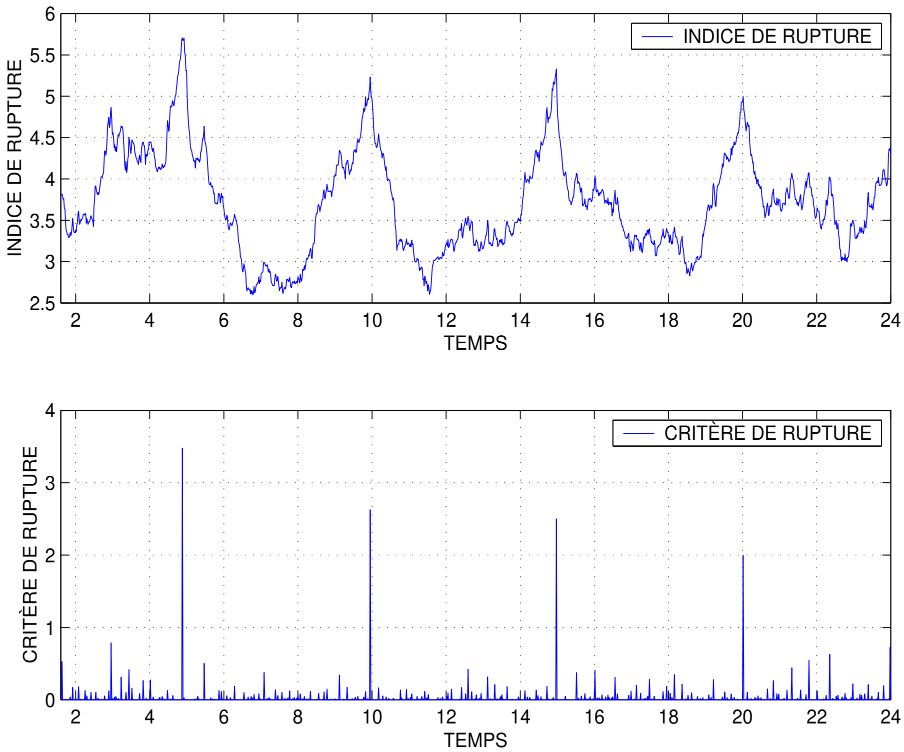

FIG. 2.3 – *Exemple d'indice de rupture (en haut) et du critere extrait (en bas).*

On aborde maintenant le cas où la distribution du signal à l'intérieur d'un segment, appartient à un ensemble fini de distributions, décrite par la famille de K densités $\{p(.|l),l = 1,\cdots,L\}$. Nous présentons dans le cas de deux distributions des segments, une extension de la méthode de base, qui utilise un indice polarisé selon le sens de la rupture. La polarisation est effectuée de manière à ce que les passages de la première distribution à la seconde correspondent aux maxima de l'indice, et les ruptures dans l'autre sens sont associées aux minima de l'indice.

Application de la méthode de base

Dans la méthode de base de détection de rupture, la famille de distributions des segments n'intervient que dans la définition de l'indice et la méthode d'extraction du critère de rupture est indépendante des distributions utilisées. Les indices de ruptures I_1 et I_2, appliqués à la famille finie de distributions ont respectivement pour expression :

$$\frac{1}{w}\ \log\ \frac{p(y_{t-w+1},\cdots,y_t|\hat{l}_{t_-})\ p(y_{t+1},\cdots,y_{t+w}|\hat{l}_{t_+})}{p(y_{t-w+1},\cdots,y_{t+w}|\hat{l}_t)}$$

$$\frac{1}{w}\left[\ \log\ \frac{p(y_{t-w+1},\cdots,y_t|\hat{l}_{t_-})}{p(y_{t-w+1},\cdots,y_t|\hat{l}_{t_+})}\ +\ \log\ \frac{p(y_{t+1},\cdots,y_{t+w}|\hat{l}_{t_+})}{p(y_{t+1},\cdots,y_{t+w}|\hat{l}_{t_-})}\ \right]$$

où :

$$\begin{cases} \hat{l}_{t_-} = \underset{l=1,\cdots,L}{\arg\max}\ p(y_{t-w+1},\cdots,y_t|l) \quad \hat{l}_{t_+} = \underset{l=1,\cdots,L}{\arg\max}\ p(y_{t+1},\cdots,y_{t+w}|l) \\ \hat{l}_t = \underset{l=1,\cdots,L}{\arg\max}\ p(y_{t-w+1},\cdots,y_{t+w}|l) \end{cases}$$

En théorie, rien n'empêche d'utiliser les indices I_1 et I_2 tels quels. Cependant, les estimateurs $\hat{l}_{t_-}$, $\hat{l}_{t_+}$ et $\hat{l}_t$ sont obtenus par une décision binaire de sélection au maximum de vraisemblance. On sait que ce procédé de décision présente des risques dans la pratique, où assez souvent on ne dispose pas des vraies distributions des segments, et ce sont des distributions estimées, plus ou moins entâchées d'erreurs, qui sont utilisées.

Une solution à ce problème consiste à utiliser la famille de mélanges générée par les L distributions, dont la famille de densités correspondante est :

$$\{p(.|\tau_1,\cdots,\tau_L) = \sum_{l=1}^{L} \tau_l\, p(.|l), \quad 0 \leq \tau_l \leq 1, \quad \sum_{l=1}^{L} \tau_l = 1\}$$

Les expressions des indices sont similaires à celles définies par les formules 2.1 et 2.2, avec le paramètre de distribution $\theta = (\tau_1,\cdots,\tau_L)$.

Extension de la méthode de base à un indice polarisé : cas d'une famille de deux distributions des segments

On suppose ici qu'il n'existe que deux types de segments, dont les distributions notées P_1 et P_2 admettent les densités respectives $p(.|1)$ et $p(.|2)$ (selon la notation introduite au début de cette section 2.2.1). Le document traité peut être une conversation entre deux locuteurs ou un document audiovisuel constitué de plages de parole et de musique. Dans le cas présent, on peut bien sûr appliquer la méthode de détection de rupture de base, soit en utilisant directement la famille de densité $\{p(.|1), p(.|2)\}$, soit à partir de la famille de densités de mélanges $\{p(.|\tau) = \tau\, p(.|1) + (1-\tau)\, p(.|2),\ \tau \in [0{,}1]\}$. La méthode qui va être décrite consiste non seulement à estimer les ruptures mais également à "préciser" leur sens, c'est-à-dire s'il s'agit d'une rupture de P_1 vers P_2 ou de P_2 vers P_1.

On peut distinguer deux sortes de ruptures : soit de P_1 vers P_2 ou de P_2 vers P_1. Il convient alors de définir un indice de rupture *polarisé* dont les maxima locaux indiquent une rupture de P_1 vers P_2 (par exemple) et les minima locaux une rupture de P_2 vers P_1. Pour ce faire, on propose une adaptation de l'indice de la divergence I_2, notée I_3, et dont l'expression est donnée ci-après :

$$I_3(t) = \frac{1}{w}\left[\ \log\ \frac{p(y_{t-w+1},\cdots,y_t|1)}{p(y_{t-w+1},\cdots,y_t|2)}\ +\ \log\ \frac{p(y_{t+1},\cdots,y_{t+w}|2)}{p(y_{t+1},\cdots,y_{t+w}|1)}\ \right] \tag{2.6}$$

On suppose que les variables $\{Y_t,\ t\ \geq\ 1\}$ sont indépendantes. L'indice $I_3(t)$ est une mesure de la différence première en t de la série du logarithme du rapport des densités $\{\log \frac{p(y_i|1)}{p(y_i|2)},\ t-w+1 \leq i \leq t+w\}$. La différence est ici calculée avec un créneau, mais d'autres fenêtres de différenciation peuvent également être utilisées.

En moyenne, l'indice I_3 a des maxima locaux aux instants de passage de P_1 vers P_2 et des minima locaux aux instants de transition de P_1 vers P_2. Pour prouver cette propriété, il suffit de démontrer que lors de l'apparition (et la progression) d'un instant rupture de P_1 vers P_2, dans la fenêtre du futur ($\{y_{t+1},\cdots,y_{t+w}\}$), comme schématisé sur la figure 2.4, la moyenne[2] de $I_3(t)$ est croissante en fonction du nombre k d'observations issues de la distribution P_2.

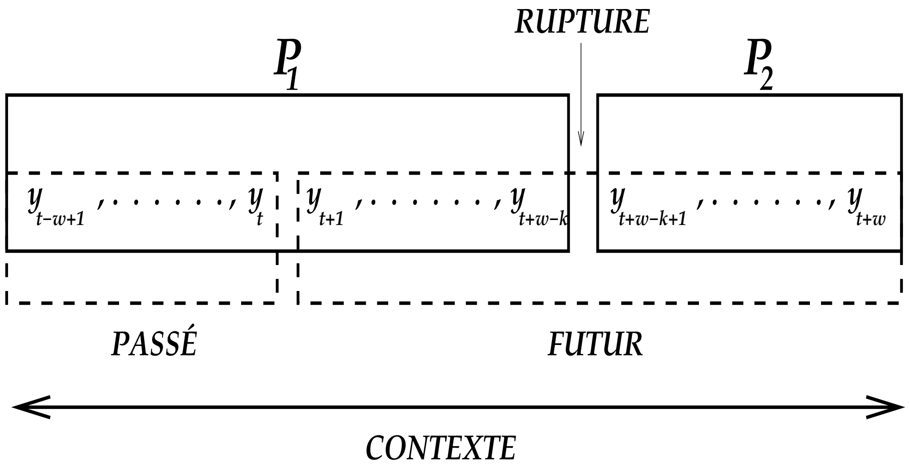

FIG. 2.4 – *Apparition d'une rupture de la distribution P_1 vers la distribution P_2 dans la fenêtre du futur. Si k est le nombre d'observations issues de P_2, l'instant de rupture est $(t+w-k)$.*

En effet, en supposant que les variables sont indépendantes, et si l'on note $\tau = \frac{k}{w}$ la proportion d'observations du futur de l'instant t, issues de la distribution P_2, on a :

$$\begin{aligned}
\mathbb{E}[I_3(t)] &= K(P_1,P_2) \,-\, (1-\tau)\,K(P_1,P_2) \,+\, \tau\,K(P_2,P_1) \\
&= \tau\,(K(P_1,P_2) \,+\, K(P_2,P_1)) \\
&= \tau\,D(P_1,P_2)
\end{aligned}$$

où $K(P,P')$ est l'information de Kullback de la distribution P par rapport à P', et $D(P,P')$ est la divergence de Kullback entre les distributions P et P'. Cela montre la propriété énoncée sur la moyenne de l'indice I_3.

Il est alors facile de constater que la moyenne $I_3(t)$ est nulle en absence de rupture dans la fenêtre du contexte, est égale à $D(P_1,P_2)$ aux instants de rupture de P_1 vers P_2, est égale à

2. Un léger abus, car il s'agit plus exactement de la moyenne de la statistique équivalente à I_3 calculée avec les variables aléatoires Y_t et non avec les observations y_t.

l'opposé de $D(P_1,P_2)$ aux instants de rupture de P_2 vers P_1, et a une pente en valeur absolue égale à $D(P_1,P_2)$.

Ce raisonnement en termes de moyenne est d'autant plus significatif que si $w \to +\infty$ et $\tau \to \tau_0$, alors $I_3(t)$ tend presque sûrement[3] vers sa moyenne.

On a déjà vu dans la méthode de détection de rupture de base, un procédé d'extraction du critère de rupture pour les maxima locaux de l'indice. On peut assez facilement en déduire l'équivalent du critère pour les minima locaux, en considérant l'opposé du critère extrait à partir de l'opposé de l'indice. Le critère final est obtenu en combinant l'extraction des maxima locaux avec celle des minima locaux. Ce critère final est positif aux instants maxima locaux, négatif aux instants minima locaux, et nul aux autres instants.

La figure 2.5 présente un exemple de l'indice de rupture I_3 et du critère extrait, qui porte sur les maxima et les minima. La décision à un instant t est prise en comparant la valeur absolue du critère en t à un seuil λ.

2.3 Expériences en segmentation parole/musique

Dans ces expériences, on traite de façon simplifiée un problème de segmentation d'un document sonore constitué de plages de parole et de musique.

2.3.1 Base de données

La base de données est constituée de morceaux de signaux de parole et de musique d'environ 4.8 secondes chacun, enregistrés sur différentes stations (une vingtaine environ, émettant à Rennes). Les signaux sont échantillonnés à la fréquence de 16 kHz. Chaque morceau de signal est ensuite découpé en trames de 16 ms (soit environ 300 trames par segment), dont chacune est paramétrisée par un vecteur de 25 Composantes Principales du Spectre (CPS)

3. C'est à dire avec une probabilité de 1 pour une suite de réalisations.

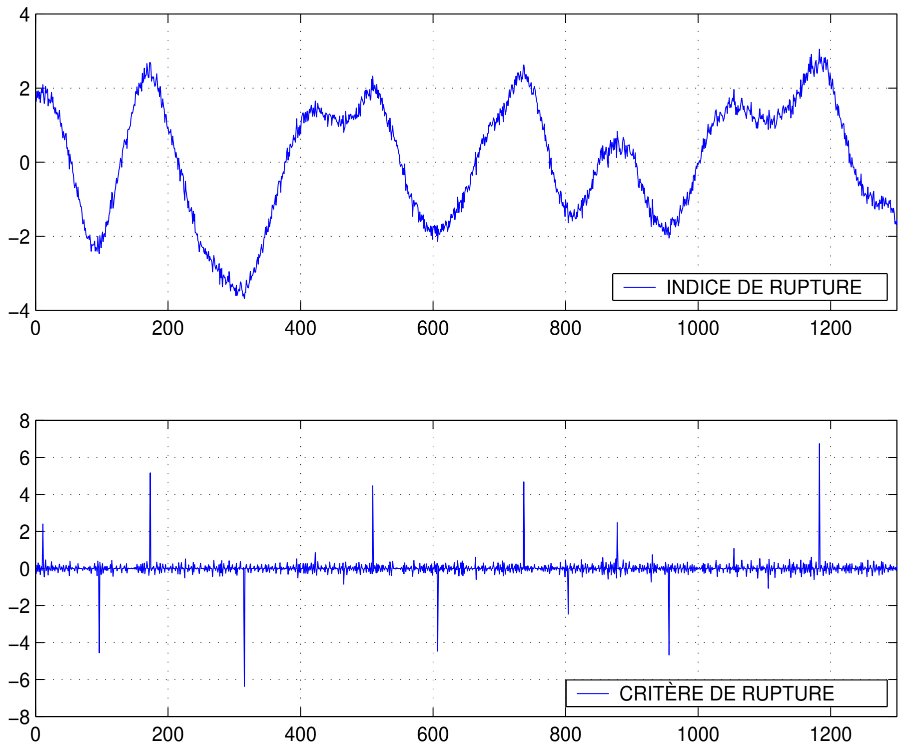

FIG. 2.5 – *Exemple d'indice de rupture (en haut) et du critère extrait (en bas), dans le cas d'un indice polarisé.*

[Seck et al., 1998]. C'est dorénavant sur le signal des vecteurs acoustiques que nous allons travailler.

2.3.2 Génération des signaux à segmenter

Afin d'évaluer l'impact de l'adéquation entre la vraie famille de distributions des segments et la famille de distributions des segments utilisée dans les algorithmes, deux signaux ont été générés : un signal à segments réels et un signal à segment simulés.

À partir de chaque segment réel, un segment simulé est réalisé en ajustant d'abord un modèle Gaussien aux vecteurs acoustiques réels, puis en générant des observations du modèle ainsi appris. Le segment simulé compte le même nombre d'observations que le segment réel associé.

Le signal à segments réels (resp. simulés) est obtenu par concaténation des segments réels (resp. simulés), en alternant ceux de parole avec ceux de musique.

Dans les expériences rapportées ici, la famille des modèles Gaussiens est utilisée.

2.3.3 Résultats des expériences

La méthode de segmentation décrite à la section 2.2 est appliquée au signal à segments simulés et au signal à segments réels, avec des modèles Gaussiens de distribution des segments, en utilisant les deux indices de rupture I_1 et I_2, et pour différentes valeurs de la taille de la fenêtre d'analyse (w).

Les performances d'une segmentation sont mesurées par le pourcentage de ruptures non détectées et le pourcentage de fausses alarmes. Une erreur de 5 trames (80 ms) sur la localisation de l'instant de rupture est tolérée. C'est à dire que chaque vrai instant de rupture r est redéfini par $r' = \arg\max_{t=r-5,\cdots,r+5} C(t)$, qui devient alors l'instant de rupture de référence dans l'évaluation des algorithmes.

Les figures 2.6 et 2.7 représentent les performances obtenues respectivement sur le signal à segments simulés et le signal à segments réels.

2.3.4 Interprétation des résultats

Les figures 2.6 et 2.7 représentent les performances obtenues par la méthode de détection de rupture décrite précédemment, sous forme de courbes DET [NIST, Eurospeech 97] qui indiquent le taux de fausse alarme en fonction du taux de non-détection lorsque le seuil de détection varie.

On observe tout d'abord sur ces figures, qu'indépendamment de l'indice de rupture

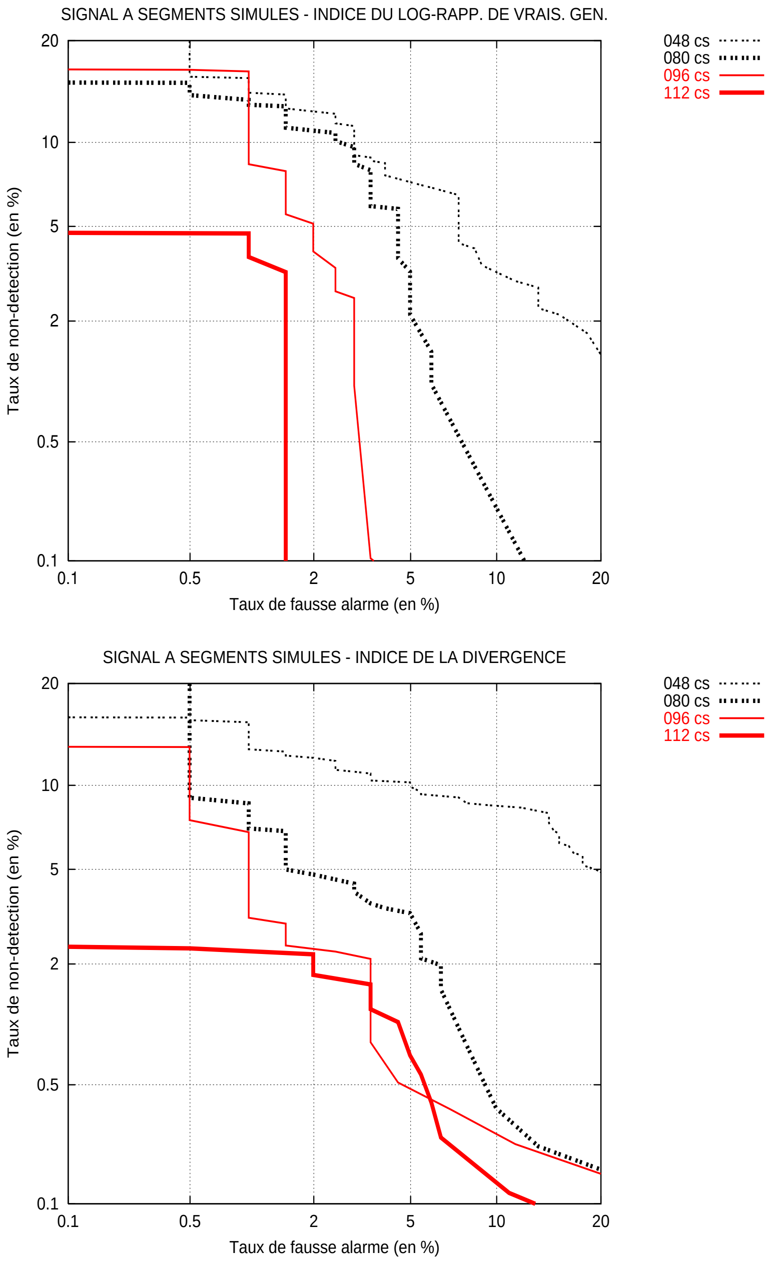

FIG. 2.6 – *Performances des deux indices de rupture sur le signal à segments simulés, pour différentes durées de la fenêtre de contexte. Les taux d'erreurs sont nuls quand la durée de la fenêtre de contexte est égale à* 1.60 *s.*

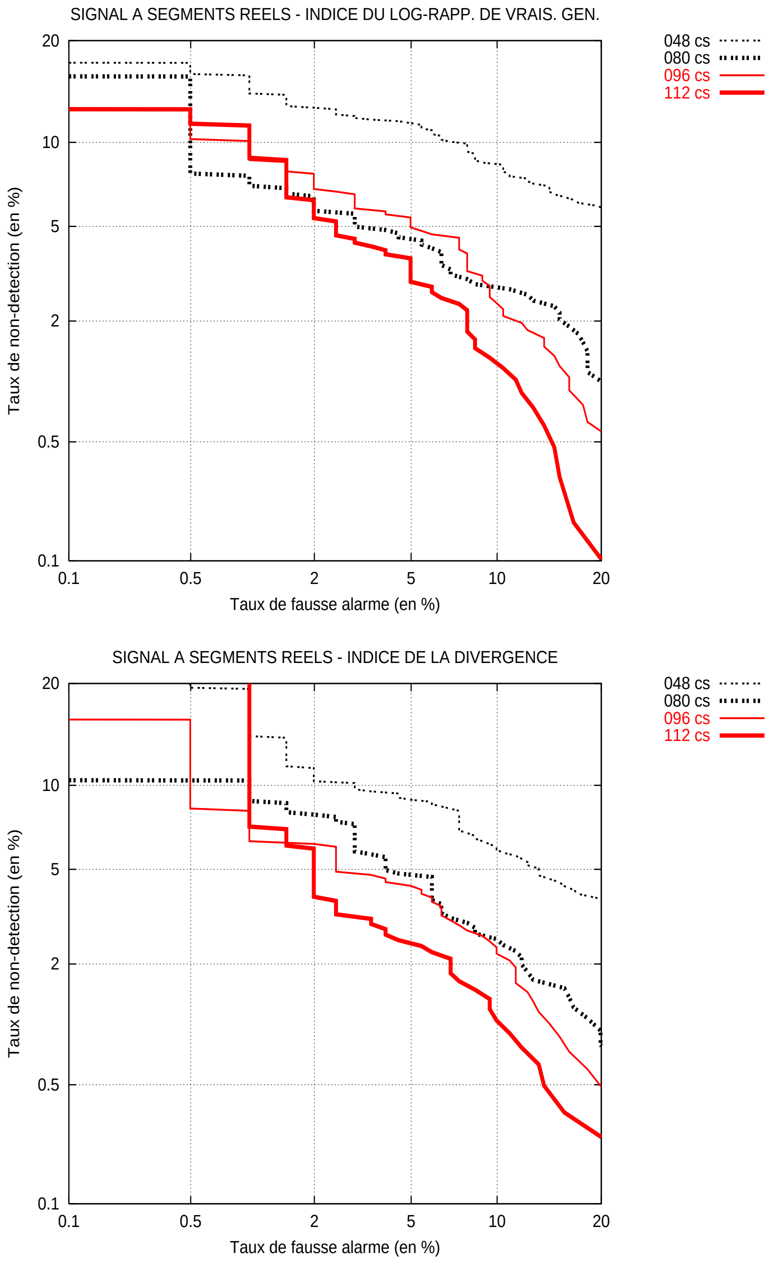

FIG. 2.7 – *Performances des deux indices de rupture sur le signal à segments réels, pour différentes durées de la fenêtre de contexte.*

utilisé, les performances sont généralement croissantes en fonction de la durée de la fenêtre de contexte. Cependant, elles semblent plafonner pour le signal à segment réels, au-delà d'une certaine durée.

De meilleures performances sont naturellement observées sur le signal à segments simulés. Pour ce dernier, les taux d'erreurs deviennent nuls à partir d'une certaine durée de la fenêtre d'analyse (1,60 s), alors que pour le signal à segments réels les performances ont tendance à se stabiliser. Cette différence de performances peut se justifier d'une part, par la cohérence entre le type de distributions (Gaussien) des segments et la famille de distributions utilisée par le procédé de segmentation dans le cas du signal des segments simulés; d'autre part, parce que sur le signal des segments réels, en plus des ruptures de type parole/musique, intervient la non-stationnarité des segments, qui génère des décisions de ruptures supplémentaires incompressibles. Néanmoins, les performances sur le signal des segments réels fournissent une première validation de l'approche proposée, avec un *Equal Error Rate* (EER) inférieur à 5% dès que la durée de la fenêtre d'analyse excède 0,80 s.

Nous n'observons pas de différence significative entre les deux indices de ruptures. Cependant, pour une configuration expérimentale avec un EER donné, on note une plus grande tendance à la non-détection pour l'indice du RVG, tandis que l'indice de la divergence s'avère plus prédisposé aux fausses alarmes.

2.4 Conclusion sur la méthode proposée

La méthode de segmentation de signaux sonores proposée repose donc sur un indice de rupture statistique suivi d'un calcul de critère de décision, sans utilisation de paramètre intermédiaire. L'indice de la divergence de Kullback, qui ne fait appel qu'à deux modèles, peut être directement adapté au cas où les distributions de probabilités des classes sont connues.

La méthode a subi un première validation par détection de ruptures sur un signal obtenu par concaténation de segments (réels et simulés) de parole et de musique en alternance. Ces expériences ne montrent pas une différence significative, en terme de performances, entre le rapport de vraisemblance généralisé et la divergence.

Par ailleurs, les performances obtenues sont plutôt rassurantes, bien que la famille de modèles utilisés soit assez simple (modèles Gaussiens). Dans la suite de nos travaux, cette méthode sera utilisée et évaluée sur des signaux audio naturels (transitions réelles entre les différents segments) et avec des modèles de mélanges de Gaussiennes.

Chapitre 3

Étude théorique des indices de rupture

Les indices de rupture I_1 et I_2 présentés à la section 2.2 ont été obtenus à partir de deux statistiques de test d'existence d'une rupture, qui sont respectivement la statistique classique du rapport de vraisemblance généralisé et une statistique basée sur la divergence de Kullback, que nous appelons *le maximum de divergence de Kullback empirique généralisée.* L'objet de ce chapitre est d'étudier ces deux statistiques et les tests basés sur ces statistiques.

On se place dans le contexte formel introduit à la section 1.1, en supposant en plus, que le signal multidimensionnel $\mathcal{Y} = \{Y_1, \cdots, Y_n, \cdots\}$ est une suite de variables aléatoires (v.a.) indépendantes. On note (toujours) y_i une observation de la v.a. Y_i. Soit $\{y_1, \cdots, y_n\}$ une séquence de n observations, on considère le test de l'hypothèse $H_0(n)$ d'absence de rupture sur la séquence, contre l'hypothèse $H_1(n)$ de présence d'une rupture.

- $H_0(n)$: *Il existe $\theta_0^\star \in \Theta$, tel que $Y_1, \cdots, Y_n$ suivent la même distribution de paramètre $\theta_0^\star$.*
- $H_1(n)$: *Il existe un instant $r^\star(n)$, avec $1 \leq r^\star(n) < n$, et deux paramètres de distribution $\theta_-^\star$ et $\theta_+^\star$, tels que $Y_1, \cdots, Y_{r^\star(n)}$ suivent la distribution de paramètre $\theta_-^\star$ et $Y_{r^\star(n)+1}, \cdots, Y_n$ suivent la distribution de paramètre $\theta_+^\star$.*

La section 3.1 introduit les deux statistiques de test d'une rupture en question, et la section 3.2 présente quelques propriétés de la statistique du maximum de divergence de Kullback empirique généralisée. La section 3.4 reprend quelques résultats généraux connus sur l'optimalité du test du rapport de vraisemblance généralisé, et ce chapitre se termine par une section 3.3 sur les lois asymptotiques des statistiques de test sous la suite d'hypothèses d'absence de rupture $H_0(n)$.

3.1 Présentation des statistiques de test étudiées

Pour le test d'existence d'une rupture, une des statistiques les plus utilisées est le rapport de vraisemblance généralisé, dont le logarithme noté $\Lambda_1(n)$ est défini par :

$$\Lambda_1(n) = \log \frac{\max\limits_{1\leq r<n} \sup\limits_{\theta_-\in\Theta} p(Y_1,\cdots,Y_r|\theta_-) \sup\limits_{\theta_+\in\Theta} p(Y_{r+1},\cdots,Y_n|\theta_+)}{\sup\limits_{\theta_0\in\Theta} p(Y_1,\cdots,Y_n|\theta_0)} \tag{3.1}$$

Relativement à un seuil λ_n, le test basé sur cette statistique est défini par sa région de rejet, dont l'expression est :

$$\left\{ \log \frac{\max\limits_{1\leq r<n} \sup\limits_{\theta_-\in\Theta} p(y_1,\cdots,y_r|\theta_-) \sup\limits_{\theta_+\in\Theta} p(y_{r+1},\cdots,y_n|\theta_+)}{\sup\limits_{\theta_0\in\Theta} p(y_1,\cdots,y_n|\theta_0)} > \lambda_n \right\}$$

L'indice $I_1(t)$ est obtenu à partir de la version instantanée du logarithme du rapport de vraisemblance généralisé (où l'instant r est fixé), notée $\Gamma_{1,n}(r)$.

$$\Gamma_{1,n}(r) = \log \frac{\sup\limits_{\theta_-\in\Theta} p(Y_1,\cdots,Y_r|\theta_-) \sup\limits_{\theta_+\in\Theta} p(Y_{r+1},\cdots,Y_n|\theta_+)}{\sup\limits_{\theta_0\in\Theta} p(Y_1,\cdots,Y_n|\theta_0)} \tag{3.2}$$

Pour obtenir $I_1(t)$ à partir de l'expresssion de $\Gamma_{1,n}(r)$, il suffit d'effectuer les correspondances $r \to t$, $n \to 2w$, puis de normaliser par $\frac{1}{w}$.

À n fixé, la distribution de la statistique $\Lambda_1(n)$ dépend de la famille de distributions utilisée, et il est en général très difficile de la relier à des lois connues. Le comportement asymptotique de $\Lambda_1(n)$ en absence de rupture (sous $H_0(n)$) a fait l'objet de plusieurs investigations, dont [Quandt, 1960, Hinkley, 1970, Feder, 1975, Bhattacharya et Brockwell, 1976], et plus récemment [Deshayes et Picard, 1983, Csörgő et Horváth, 1997].

Dans l'expression de $\Lambda_1(n)$, le maximum de vraisemblance sur l'ensemble des n variables qui apparaît en dénominateur, complique l'interprétation de cette statistique en présence de rupture, car on estime un seul paramètre de distribution sur des v.a. hétérogènes, n'ayant pas la même distribution. Ainsi, la loi asymptotique de $\Lambda_1(n)$ en présence de rupture dépend en général de manière très compliquée des paramètres avant et après rupture ($\theta_-^\star$ et $\theta_+^\star$), comme le montrent les travaux de [Csörgő et Horváth, 1997] où les distributions sont de type exponentiel (lois Gaussiennes, Exponentielles, de Poisson, etc...).

Pour les raisons qui viennent d'être évoquées, nous proposons une autre statistique de test d'existence d'une rupture, qui est une mesure empirique du maximum sur tous les instants r de la divergence de Kullback entre les distributions avant et après r. Cette statistique est notée $\Lambda_2(n)$, et a pour expression :

$$\Lambda_2(n) = \max_{1 \leq r \leq n-1} n \left[\frac{1}{r} \log \frac{p(Y_1, \cdots, Y_r | \hat{\theta}_{r_-})}{p(Y_1, \cdots, Y_r | \hat{\theta}_{r_+})} + \frac{1}{n-r} \log \frac{p(Y_{r+1}, \cdots, Y_n | \hat{\theta}_{r_+})}{p(Y_{r+1}, \cdots, Y_n | \hat{\theta}_{r_-})} \right] \tag{3.3}$$

où :

$$\hat{\theta}_{r_-} = \arg\max_{\theta \in \Theta} p(Y_1, \cdots, Y_r | \theta) \quad \text{et} \quad \hat{\theta}_{r_+} = \arg\max_{\theta \in \Theta} p(Y_{r+1}, \cdots, Y_n | \theta)$$

On suppose implicitement que la famille de densités est telle que les estimateurs du maximum de vraisemblance (EMV) soient définis pour toutes valeurs d'une séquence finie de v.a.

Soit $\Gamma_{2,n}(r)$ l'expression en argument du max dans la formule 3.3 définissant $\Lambda_2(n)$:

$$\begin{aligned}\Gamma_{2,n}(r) &= n\left[\frac{1}{r}\,\log\,\frac{p(Y_1,\cdots,Y_r|\hat\theta_{r_-})}{p(Y_1,\cdots,Y_r|\hat\theta_{r_+})} + \frac{1}{n-r}\,\log\,\frac{p(Y_{r+1},\cdots,Y_n|\hat\theta_{r_+})}{p(Y_{r+1},\cdots,Y_n|\hat\theta_{r_-})}\right]\\ &= n\left[\frac{1}{r}\sum_{i=1}^{r}\log\,\frac{p(Y_i|\hat\theta_{r_-})}{p(Y_i|\hat\theta_{r_+})} + \frac{1}{n-r}\sum_{i=r+1}^{n}\log\,\frac{p(Y_i|\hat\theta_{r_+})}{p(Y_i|\hat\theta_{r_-})}\right] \qquad (3.4)\end{aligned}$$

Nous appelons $\Gamma_{2,n}(r)$ *la divergence de Kullback empirique généralisée* à l'instant r basée sur $\{Y_1,\cdots,Y_n\}$: elle est *empirique* dans le sens où elle est fonction de variables aléatoires, et elle est *généralisée* parce qu'elle utilise des paramètres de distributions estimés au maximum de vraisemblance à la place de paramètres fixes, par analogie au rapport de vraisemblance généralisé. Nous avons alors choisi d'appeler $\Lambda_2(n)$ *le maximum de divergence de Kullback empirique généralisée.*

L'indice de rupture $I_2(t)$ est obtenu à partir de $\Gamma_{2,n}(r)$, en effectuant les mêmes correspondances permettant de passer de $\Gamma_{1,n}(r)$ à $I_1(t)$.

3.2 Propriétés de la divergence de Kullback empirique généralisée dans le cas d'une famille de densités de type exponentiel

L'objet de cette section est d'énoncer quelques propriétés illustratives de la divergence de Kullback empirique généralisée à l'instant r, $\Gamma_{2,n}(r)$, dans le cas d'une famille de densités de type exponentiel. On dit qu'une famille de densités $\{p(.|\theta),\theta \in \Theta \subset \mathbb{R}^m\}$ est de type exponentiel, si elle est de la forme[1] :

$$p(y|\theta) \;=\; \exp\left[\theta^T S(y) + U(y) - A(\theta)\right]\;\mathbf{1}_{\{y\in C\}} \qquad (3.5)$$

1. Dans cette formulation, la paramétrisation est sous sa forme naturelle.

Le type exponentiel regroupe, par exemple, les familles de lois Gaussiennes, Exponentielles, de Poisson, etc...

Pour une famille de distributions de type exponentiel, l'hypothèse d'existence et d'unicité des estimateurs du maximum de vraisemblance revient à supposer que le gradient de A, noté A', admet une réciproque (ou inverse), notée $(invA')$. Les estimateurs du maximum de vraisemblance, $\hat{\theta}_{r_-}$ et $\hat{\theta}_{r_+}$, ont alors pour expression :

$$\hat{\theta}_{r_-} = (invA')(\bar{S}_{r_-}) \quad \text{et} \quad \hat{\theta}_{r_+} = (invA')(\bar{S}_{r_+})$$

où :

$$\bar{S}_{r_-} = \frac{1}{r}\sum_{i=1}^{r} S(Y_i) \quad \text{et} \quad \bar{S}_{r_+} = \frac{1}{n-r}\sum_{i=r+1}^{n} S(Y_i)$$

Les estimateurs du maximum de vraisemblance, $\hat{\theta}_{r_-}$ et $\hat{\theta}_{r_+}$, s'expriment ainsi comme une fonction $(invA')$ des moyennes empiriques $\bar{S}_{r_-}$ et $\bar{S}_{r_+}$. Une relation similaire relie un paramètre de distribution θ à la moyenne de $S(Y)$, quand Y suit la distribution de paramètre θ :

$$\theta = (invA')(\mathbb{E}_\theta\left[S(Y)\right])$$

On peut voir assez facilement d'après la forme des densités donnée par la formule 3.5, que la divergence de Kullback, $D(\theta_1,\theta_2)$, entre deux distributions de paramètres θ_1 et θ_2, a pour expression :

$$\begin{aligned} D(\theta_1,\theta_2) &= \{\theta_1 - \theta_2\}^T \{\mathbb{E}_{\theta_1}\left[S(Y)\right] - \mathbb{E}_{\theta_2}\left[S(Y)\right]\} \\ &= \{\theta_1 - \theta_2\}^T \left\{A'(\theta_1) - A'(\theta_2)\right\} \qquad (3.6) \end{aligned}$$

Un calcul immédiat permet aussi de voir que :

$$\Gamma_{2,n}(r) = n\left\{\hat{\theta}_{r-} - \hat{\theta}_{r+}\right\}^T \left\{A'(\hat{\theta}_{r-}) - A'(\hat{\theta}_{r+})\right\} \qquad (3.7)$$

Ainsi, les formules 3.6 et 3.7 montrent que $\frac{1}{n}\Gamma_{2,n}(r)$ est exactement la divergence de Kullback entre les distributions de paramètres $\hat{\theta}_{r_-}$ et $\hat{\theta}_{r_+}$.

On montre de la même manière que le résultat sur la moyenne de l'indice $I_3(t)$ (section 2.2.1), que si les paramètres avant et après rupture $\theta_-^\star$ et $\theta_-+^\star$ sont fixés, alors la moyenne de la statistique $\frac{1}{n}\Gamma_{2,n}(r)$ correspondante atteint son maximum à l'instant de rupture.

Considérons maintenant une suite d'expériences statistiques indicées par n, correspondant à une apparition de rupture en un instant $r_n^\star$, où $0 < r_n \leq r_n^\star < n$, telles que quand $n \to +\infty$, $r_n \to +\infty$, $(r_n^\star - r_n) \to +\infty$, $(n - r_n^\star) \to +\infty$, et la position de l'instant de rupture $\frac{n-r_n^\star}{n-r_n}$ tend vers une proportion $\tau \in]0,1[$.

Les statistiques $\bar{S}_{r_-}$ et $\bar{S}_{r_+}$ sont des combinaisons linéaires de moyennes de v.a. i.i.d, donc sous les bonnes hypothèses de la Loi des Grands nombres, leur convergence presque sûre est assurée. Ces résultats peuvent être en mesure de faciliter l'étude des propriétés asymptotiques des statistiques $\Gamma_{2,n}(r_n)$ et $\Lambda_2(n)$, lorsque n tend vers $+\infty$.

Pour des familles de distributions de type non-exponentiel, nous n'avons pas, pour le moment, de résultats théoriques exacts, concernant les propriétes asymptotiques de $\Gamma_{2,n}(r_n)$ et $\Lambda_2(n)$.

3.3 Comportement asymptotique des statistiques de test en absence de rupture

Cette section est consacrée à l'étude du comportement asymptotique sous $H_0(n)$, des statistiques définies sur la base des rapports de vraisemblance généralisés $\{\Gamma_{1,n}(r),\ 1 \leq r < n\}$ ou des divergences de Kullback empiriques généralisées $\{\Gamma_{2,n}(r),\ 1 \leq r < n\}$. Ces statistiques sont obtenues en pondérant en fonction de l'instant r, $\Gamma_{1,n}(r)$ ou $\Gamma_{2,n}(r)$ avant de prendre le maximum sur tous les instants r. Nous verrons plutard le sens de cette pondération.

Nous commençons par présenter des résultats de convergence en loi connus, relatifs aux statistiques définies à partir de $\Gamma_{1,n}$. Ensuite, nous énonçons un résultat analogue sur le comportement asymptotique des statistiques construites à partir de $\Gamma_{2,n}$.

3.3.1 Comportement asymptotique du rapport de vraisemblance généralisé

Des études du comportement asymptotique de la statistique $\Lambda_1(n)$, sous la suite l'hypothèses d'absence de rupture $H_0(n)$, sont présentées dans [Deshayes et Picard, 1986]. Sous $H_0(n)$ et certaines hypothèses de régularité sur la famille de distributions, Deshayes et Picard montrent que la statistique

$$\Lambda_1'(n) = \max_{1\leq r\leq n-1} \psi\left(\frac{r}{n}\right) \Gamma_{1,n}(r) \quad \text{converge en loi vers} \quad \sup_{0\leq\tau\leq 1} \frac{\psi(\tau)}{2\tau(1-\tau)} B^{(m)}(\tau) \tag{3.8}$$

Dans cette relation, $B^{(m)}$ est la somme des carrés de m ponts Browniens indépendants réels, $\{B_i,\, i=1,\cdots,m\}$, définis sur [0,1].

$$B^{(m)}(\tau) \;=\; \sum_{i=1}^{m} B_i^2(\tau) \tag{3.9}$$

Un pont Brownien B est défini à l'aide d'un mouvement Brownien W par la relation

$$B(\tau) \;=\; W(\tau) \;-\; \tau W(1) \qquad \text{pour} \qquad \tau \in [0{,}1] \tag{3.10}$$

Quant à la fonction ψ, elle est définie sur [0,1], positive et telle que

$$\int_0^1 \left[\frac{\psi(\tau)}{\tau(1-\tau)}\right]^2 d\tau < +\infty. \tag{3.11}$$

Ce résultat met en évidence la nécessité de pondérer la statistique $\Gamma_{1,n}(r)$ avant de prendre le maximum sur tous les instants r. Dans le cas où $\psi(\tau) = 2\tau(1-\tau)$, la distribution limite du résultat 3.8, $\sup_{0\leq\tau\leq 1} B^{(m)}(\tau)$, est connue [Shorack et A., 1986, Kiefer, 1959a, Kiefer, 1959b].

Les travaux présentés dans [Csörgő et Horváth, 1997], qui supposent d'autres conditions énoncées en annexe A.1, retrouvent un résultat de convergence similaire. Pour énoncer le résultat de Csörgő et Horváth, on introduit quelques notations. On note alors $V_{1,n}$ le processus à temps continu défini sur [0,1] par :

$$V_{1,n}(\tau) = \begin{cases} 0 & \text{si} \quad 0 \leq \tau < \frac{1}{n+1} \\ 2\Gamma_{1,n+1}([(n+1)\tau]) & \text{si} \quad \frac{1}{n+1} \leq \tau < \frac{n}{n+1} \\ 0 & \text{si} \quad \frac{n}{n+1} \leq \tau \leq 1 \end{cases} \tag{3.12}$$

où on utilise la notation usuelle $[x]$ pour désigner la partie entière de x.

$V_{1,n}$ est la version continue de $\Gamma_{1,n+1}$.

On dispose maintenant de tous les éléments pour énoncer le résultat de Csörgő et Horváth.

Théorème 3 (Csörgő et Horváth)
On se place sous la suite d'hypothèses $H_0(n)$ et les conditions (C1) à (C9) en annexe A.1. Soit q une fonction positive sur]0,1[, croissante au voisinage de 0 et décroissante au voisinage de 1. Alors :

$$\sup_{0\leq\tau\leq 1} \frac{\tau(1-\tau)V_{1,n}(\tau)}{q^2(\tau)} \qquad \textit{converge en loi vers} \qquad \sup_{0\leq\tau\leq 1} \frac{B^{(m)}(\tau)}{q^2(\tau)} \tag{3.13}$$

si et seulement si, il existe une constante $c > 0$ pour laquelle :

$$I_{0,1}(q,c) = \int_0^1 \frac{1}{\tau(1-\tau)} \exp\left(-\frac{cq^2(\tau)}{\tau(1-\tau)}\right) d\tau \quad < \quad +\infty \tag{3.14}$$

Nous allons énonçer dans la prochaine section un résultat de convergence similaire pour le processus $\Gamma_{2,n}$.

3.3.2 Comportement asymptotique de la divergence de Kullback empirique généralisée

Pour l'étude du comportement asymptotique sous $H_0(n)$ des statistiques définies à partir du processus $\{\Gamma_{2,n}(r),\ 1 \leq r < n\}$, en pondérant avant de prendre le maximum sur tous les instants r, nous allons suivre [Csörgő et Horváth, 1997], dont l'approche utilise explicitement la valeur des estimateurs du maximum de vraisemblance, comme le fait $\Gamma_{2,n}$. De manière analogue à $V_{1,n}$, nous définissons alors à l'aide de $\Gamma_{2,n}$, le processus à temps continu $V_{2,n}$ par :

$$V_{2,n}(\tau) = \begin{cases} 0 & \text{si} \quad 0 \leq \tau < \frac{1}{n+1} \\ \Gamma_{2,n+1}([(n+1)\tau]) & \text{si} \quad \frac{1}{n+1} \leq \tau < \frac{n}{n+1} \\ 0 & \text{si} \quad \frac{n}{n+1} \leq \tau \leq 1 \end{cases} \tag{3.15}$$

Nous montrons alors sous les mêmes conditions le résultat analogue au théorème 3 de Csörgő et Horvàth.

Théorème 4

Sous les mêmes hypothèses qu'au théorème 3 :

$$\sup_{0\leq\tau\leq 1} \frac{\tau^2(1-\tau)^2 V_{2,n}(\tau)}{q^2(\tau)} \qquad \textit{converge en loi vers} \qquad \sup_{0\leq\tau\leq 1} \frac{B^{(m)}(\tau)}{q^2(\tau)} \tag{3.16}$$

si et seulement si, il existe une constante $c > 0$ *pour laquelle* $I_{0,1}(q,c) < +\infty$.

La preuve de ce théorème est un peu longue et est présentée en annexe A. On attire l'attention du lecteur sur le fait que la seule différence entre les deux résultats des théorèmes 3 et 4 est la correspondance $\tau(1-\tau) \longleftrightarrow \tau^2(1-\tau)^2$.

3.3.3 Conclusion sur le comportement asymptotique des statistiques de test d'une rupture

Cette section a été l'objet de l'étude du comportement asymptotique en absence de rupture, des statistiques définies à partir des rapports de vraisemblance généralisés instantanés

$\{\Gamma_{1,n}(r),\ 1 \leq r < n\}$, ou des divergences de Kullback empiriques généralisées instantanées $\{\Gamma_{2,n}(r),\ 1 \leq r < n\}$, en pondérant en fonction de l'instant r avant de prendre de maximum sur tous les instants. Les lois asymptotiques des statistiques relatives à $\Gamma_{1,n}$ sont connues sous certaines conditions de régularité sur la famille de distributions et sur la fonction de pondération. Sous des conditions très similaires, des propriétés asymptotiques analogues ont été énoncées pour les statistiques obtenues à partir de $\Gamma_{2,n}$.

Les statistiques de test qui sont définies ici, utilisent la vraie famille de densités de la suite de v.a. $\{p(.|\theta), \theta \in \Theta \subset \mathbb{R}^m\}$. Certaines autres familles de fonctions, répondant aux mêmes conditions générales de régularités, peuvent conduire à des statistiques ayant les mêmes lois asymptotiques en absence de rupture, que $\Lambda'_1(n)$ ou $\Lambda'_2(n)$. C'est le cas des statistiques de test de changement de moyenne en utilisant la famille de densités Gaussiennes de variance égale à l'identité. Ainsi, l'utilisation de la vraie famille de distributions n'est pas toujours importante en absence de rupture, pour aboutir aux lois asymptotiques des statistiques utilisant la vraie famille de distributions.

3.4 Résultats d'optimalité du test du rapport de vraisemblance généralisé

Cette section présente quelques propriétés d'optimalité connues du test du rapport de vraisemblance généralisé pour la détection d'une rupture. Ces résultats sont énoncés plus en détail dans [Deshayes et Picard, 1986, Genot-Catalot et Picard, 1993]. Plusieurs cas de figures sont envisagés selon le niveau de connaissance *a priori* sur l'instant de rupture et les paramètres de distribution avant et après rupture.

3.4.1 Paramètres avant et après rupture et instant de test fixés

Dans le cas le plus simple où l'instant de test de la rupture, le paramètre en absence rupture et les paramètres avant et après rupture sont tous fixés, on est dans la situation d'un test entre deux hypothèses simples, et le théorème de Neyman-Pearson garantit l'optimalité du test du rapport de vraisemblance.

3.4.2 Cas plus généraux

Absence de test optimal à n fixé

Dans le cas général où l'instant de test de la rupture, le paramètre en absence de rupture et les paramètres avant et après rupture ne sont pas fixés, on a vu qu'on les remplace par leurs estimateurs du maximum de vraisemblance dans l'expression du rapport de vraisemblance.

À nombre d'observations n fixé, dès que l'instant où on teste la rupture est inconnu, il n'y a pas de test optimal au sens Uniformément Plus Puissant, pour des cas très simples. Deshayes et Picard ont donné l'exemple du test d'un changement de moyenne, dans une suite de v.a. Gaussiennes de variance 1; les moyennes avant et après rupture étant connues.

Cela incite à étudier les propriétés d'optimalité du test du rapport de vraisemblance généralisé, dans un cadre asymptotique.

Absence de test localement optimal

On a vu à la section 2.4 de la partie I, qu'il existe deux cadres asymptotiques de comparaison des tests : l'asymptotique local et l'asymptotique non-local. Dans le cas simple Gaussien qu'on vient de citer, il n'existe pas non plus de test Localement Asymptotiquement Plus Puissant.

Optimalité non-locale dans le cas de distributions Gaussiennes

La proposition qui suit, due à Deshayes et Picard, énonce l'optimalité asymptotique non-locale du test du rvg dans le cas de familles de distributions Gaussiennes, selon plusieurs niveaux de connaissance *a priori* sur les paramètres.

Proposition 1

Quand le log-niveau α est fixé, le test du rapport de vraisemblance, basé sur la statistique $\Lambda_1(n)$, donne la meilleure log-erreur $\beta(\theta_-,\tau,\theta_+)$, à tout point de l'hypothèse de présence de rupture (θ_-,τ,θ_+), dans les cas de figures suivants :

- *parmi tous les tests invariants par translation, pour le model $\mathcal{N}(\mu,1)$: changement de moyenne dans une suite de variables Gaussiennes de variance 1 connue.*
- *parmi tous les tests invariants par homothétie-translation, pour le model $\mathcal{N}(\mu,\sigma^2)$: changement de moyenne dans une suite de variables Gaussiennes de variance inconnue.*
- *parmi tous les tests invariants pour le model $\mathcal{N}(0,\sigma^2)$: changement de variance dans une suite de variables Gaussiennes de moyenne connue.*

Pour d'autres familles de distributions, il n'existe pas de résultat d'optimalité asymptotique non-locale. Actuellement, les lois limites en présence de rupture ne sont obtenues que pour des distributions appartenant à la famille exponentielle [Csörgő et Horváth, 1997].

3.5 Conclusion

Ce chapitre a présenté une étude de tests d'existence d'une rupture, basés sur deux statistiques de test : le rapport de vraisemblance généralisé usuel et une statistique alternative introduite dans ce travail, le maximum de divergence de Kullback empirique généralisée. Quand l'instant de test de la rupture est fixé, des propriétés de cette dernière statistique par rapport à la divergence de Kullback entre les estimateurs du maximum de vraisemblance des paramètres avant et après l'instant courant, on été établies pour des distributions de type exponentiel. Dans le cas où l'instant de test de la rupture n'est pas fixé, la loi asymptotique en absence de rupture de cette statistique alternative a également été trouvée. Elle s'avère d'ailleurs assez proche de la loi asymptotique (en absence de rupture) connue du rapport de vraisemblance généralisé.

Pour des distributions Gaussiennes, on sait que le test du rapport de vraisemblance généralisé est non-localement asymptotiquement optimal. Il apparait alors naturel de s'interroger sur l'optimalité du test basé sur le maximum de divergence de Kullback empirique généralisée, pour cette famille de distributions relativement simple. Des efforts supplémentaires seront consentis sur l'étude des propriétés (asymptotiques) de cette statistique en présence de rupture, en liaison avec la divergence de Kullback entre les distributions avant et après rupture.

Troisième partie

DÉTECTION ET SUIVI D'UNE CLASSE DE SONS DANS UN DOCUMENT

Chapitre 1

Présentation des tâches

Dans cette partie, on considére un document sonore D, constitué de deux classes de sons : une classe cible notée X, qui correspond à une classe de sons recherchée, et une classe non-cible notée $\overline{X}$, qui regroupe tous les sons qui n'appartiennent pas à la classe cible. Il peut s'agir d'un document audiovisuel contenant de la parole et d'autres types de sons comme de la musique, ou d'une conversation entre une personne cible et d'autres locuteurs, etc... Il est important de préciser que le document peut être exclusivement constitué, soit de la classe de sons cible, soit de la classe non-cible.

Dans le but d'extraire une information descriptive du document, différentes tâches peuvent être accomplies. On s'intéresse à deux d'entre elles, qui répondent à deux questions que l'on peut se poser sur ce type de document :

1. La classe de sons cible est-elle présente dans le document ?
2. À quel(s) endroit(s) du document est-elle présente ?

La première question pose un problème de *détection* de la classe cible et est abordée dans la section 2. La seconde question correspond au problème de *suivi* de la classe cible le long du document et est traitée dans la section 3.

Pour ces deux types de tâches, le document est paramétrisé et $\{y_1, \cdots, y_n\}$ dénotent les vecteurs de paramètres acoustiques extraits du document.

Les algorithmes présentés sont basés sur des méthodes statistiques et utilisent deux distributions de paramètres acoustiques : la distribution de la classe cible notée P_1, et la distribution de la classe non-cible notée P_0. Dans la pratique, comme dans toutes les expériences qui sont présentées, les distributions exactes sont inconnues. On utilise alors des approximations des vraies distributions, en ajustant des modèles sur des données d'apprentissage.

Dans ce travail, on utilise des Modèles de Mélange de Gaussiennes (MMG) pour modéliser les distributions de probabilité des classes. Ces modèles, ainsi que leurs extensions aux Modèles de Markov Cachés (MMC), constituent l'etat de l'art actuel, dans le domaine du Traitement Automatique de la Parole et du Son.

Une des particularités de l'approche présentée ci-après est de s'appuyer sur un modèle de *mélange* des distributions de probabilité entre la classe cible et la classe non-cible, pour caractériser la distribution de probabilité des paramètres acoustiques au sein d'un document mixte. Cette approche est testée et comparée à d'autres approches plus conventionnelles, d'une part dans le cadre de la détection parole/musique (corpus AIM du projet DiVAN), d'autre part dans le contexte des évaluations NIST en détection et suivi du locuteur (corpus SWITCHBOARD, Consortium ELISA).

Chapitre 2

Algorithmes de détection d'une classe de sons dans un document

La tâche de détection d'une classe de sons dans un document consiste à dire si une classe de sons cible est présente dans un document D, au vu de la séquence de n vecteurs de paramètres acoustiques extraits du document $\{y_1, \cdots, y_n\}$, connaissant la distribution des vecteurs acoustiques cibles, P_1, et celle non-cible, P_0. On suppose que P_0 et P_1 admettent des densités qui sont respectivement notées p_0 et p_1.

En adoptant une approche statistique, nous proposons de placer ce problème dans le cadre du test entre deux hypothèses :

- $H_0(n)$: $\{y_1, \cdots, y_n\}$ *est entièrement générée par une v.a. de loi* P_0.
- $H_1(n)$: $\{y_1, \cdots, y_n\}$ *est en partie généré par une v.a.de loi* P_1.

On considère que $\{y_1, \cdots, y_n, \cdots\}$ sont des observations d'une suite de variables aléatoires $\{Y_1, \cdots, Y_n, \cdots\}$ indépendantes et identiquement distribuées. La distribution P_0 supposée être connue, $H_0(n)$ est une hypothèse simple. Cependant, relativement à l'hypothèse $H_1(n)$, deux cas de figures peuvent être envisagés, selon que l'on suppose qu'il puisse y avoir présence conjointe des deux distributions, cible et non-cible, ou pas.

2.1 Document exclusivement cible ou non-cible

Dans un premier cas, on exclut la situation où l'échantillon compte aussi bien des observations de la distribution cible que celles de la distribution non-cible. L'hypothèse H_1 devient plus précisément l'hypothèse simple "$\{y_1, \cdots, y_n\}$ *est entièrement généré par une v.a. de distribution* P_1". Il s'agit dans ce cas d'un test entre deux hypothèses simples, pour lequel le test du rapport de vraisemblance est optimal (au sens Uniformément Plus Puissant, d'après le théorème connu de Neyman-Pearson). Ce test est basé sur la statistique du rapport de vraisemblance, dont le logarithme noté $s_0(n)$ est donné par :

$$s_0(n) = \log \frac{p_1(y_1, \cdots, y_n)}{p_0(y_1, \cdots, y_n)} \tag{2.1}$$

La région de rejet de ce test, pour un seuil $\lambda_0(n)$ est

$$\{s_0(n) \ > \ \lambda_0(n)\}$$

2.2 Document éventuellement mixte

Dans un second cas, on suppose qu'il peut y avoir une présence conjointe des distributions cible et non-cible dans la séquence d'observations. La distribution sous $H_1(n)$ n'est alors pas entièrement définie par la seule connaissance de P_1 et de P_0, car elle dépend également de la manière dont les distributions cible et non-cible coexistent. Par exemple, on peut se demander quelles sont les proportions de cible et de non-cible? Les termes de cette coexistence peuvent être associés à un paramètre supplémentaire dans la distribution sous $H_1(n)$. La section suivante présente un modèle de présence conjointe des distributions cible et non-cible.

2.2.1 Modèle de présence conjointe des deux distributions

Pour prendre en compte la présence conjointe des distributions cible et non-cible, on propose une approche basée sur un modèle de mélange classique de ces deux distributions.

D'autres types de modèles, comme les Modèles de Markov Cachés ou les Modèles Segmentaux, peuvent également être utilisés. Le modèle de mélange est choisi pour sa simplicité. La densité du modèle de mélange est déterminée par le poids de la distribution cible que l'on note θ, et qui appartient à [0,1]. Cette densité, notée p_θ, a pour expression en y

$$p_\theta(y) = \theta * p_1(y) + (1-\theta) * p_0(y) \quad , \quad \theta \in [0,1] \tag{2.2}$$

Dans la formule 2.2, θ est la probabilité *a priori* qu'une observation provienne de la distribution cible P_1.

Dans le contexte du modèle 2.2, si $\theta_n^\star$ dénote le vrai paramètre de distribution des observations $\{y_1,\cdots,y_n\}$, les hypothèses de test $H_0(n)$ et $H_1(n)$, énoncées au début de ce chapitre, peuvent être reformulées plus précisément :

- $H_0(n): \quad \theta_n^\star = 0$
- $H_1(n): \quad 0 < \theta_n^\star \leq 1$

2.2.2 Proposition de tests statistiques

Test du rapport de vraisemblance de P_0 contre P_1

La statistique $s_0(n)$ (formule 2.1), bien que plus approriée au cas où les observations sont exclusivement cibles ou non-cibles, possède des propriétés intéressantes pour la présente situation où, sous $H_1(n)$, les observations sont issues d'un mélange des distributions P_0 et P_1. Soit $S_0(n)$ définie par :

$$S_0(n) = \sum_{i=1}^{n} \log \frac{p_1(Y_i)}{p_0(Y_i)} \tag{2.3}$$

Si $\tau(n)$ dénote la proportion de v.a. suivant la distribution cible P_1,

$$\mathbb{E}\left[\frac{1}{n} S_0(n)\right] = -K(P_0,P_1) + \tau(n)\, D(P_1,P_0) \tag{2.4}$$

où $K(P_0,P_1)$ est l'information de Kullback de P_0 par raport à P_1 et $D(P_1,P_0) = K(P_1,P_0) + K(P_0,P_1)$, est la divergence de Kullback entre P_0 et P_1. Comme on suppose que les distributions

P_1 et P_0 sont différentes, $D(P_1,P_0)$ est strictement positive, la statistique $S_0(n)$ est en moyenne strictement croissante par rapport à $\tau(n)$, et dès que $\tau(n) > 0$ (sous $H_1(n)$) on a :

$$\mathbb{E}_{H_0(n)}[S_0(n)] \; < \; \mathbb{E}_{H_1(n)}[S_0(n)] \tag{2.5}$$

Par ailleurs, on peut voir en utilisant la loi des grands nombres que si $\tau(n) \to \tau$, alors :

$$S_0(n) \qquad \text{converge presque surement vers} \qquad [-K(P_0,P_1) \; + \; \tau\, D(P_1,P_0)] \tag{2.6}$$

Cette propriété est cohérente avec l'utilisation de $S_0(n)$ pour la détection d'une classe de sons dans un document mixte. A cet effet, on propose le test dont la fonction de rejet notée $\varphi_0(n)$ est définie par :

$$\varphi_0(n) = \mathbf{1}_{\{s_0(n) > \lambda_0(n)\}} \tag{2.7}$$

Pour un niveau α, la constante $\lambda_0(n)$ est choisie telle que :

$$\mathbb{E}_{H_0(n)}[\Phi_0(n)] = \alpha$$

où est $\Phi_0(n)$ la statistique portant sur les v.a.

$$\Phi_0(n) = \mathbf{1}_{\{S_0(n) > \lambda_0(n)\}}$$

Test de l'EMV du poids de P_1 dans le modèle de mélange $\{P_1,P_0\}$

Dans notre situation, le vrai poids de la distribution cible est inconnu. Une estimation de ce poids sur une séquence d'observations mélange, mesure la proportion d'observations provenant de la distribution cible.

On propose alors comme statistique de détection de la classe cible, l'estimateur du maximum de vraisemblance du poids de la distribution cible, noté $\hat{\theta}_n$ ou $s_1(n)$:

$$s_1(n) = \hat{\theta}_n = \arg\max_{\theta \in [0,1]} \; p_\theta(y_1, \cdots, y_n) \tag{2.8}$$

Le test proposé, basé sur $s_1(n)$, est donné pour un seuil $\lambda_1(n)$ par sa fonction de rejet notée $\varphi_1(n)$, définie par :

$$\varphi_1(n) = \mathbf{1}_{\{s_1(n)>\lambda_1(n)\}} \tag{2.9}$$

La constante $\lambda_1(n)$ peut être choisie en fonction d'un niveau α requis.

L'estimateur $\hat{\theta}_n$ peut être facilement calculé au moyen de l'algorithme EM pour les modèles de mélange. Et notre cas est d'autant plus simple qu'il n'y a qu'un seul poids à estimer. D'autres estimateurs peuvent également être utilisés, mais l'e.m.v. est préféré, compte tenu des outils théoriques existants à son sujet.

Test du rapport de vraisemblance généralisé

Une autre proposition de statistique de détection, plus naturelle du point de vue de la théorie des tests statistiques, est la statistique du rapport de vraisemblance généralisé, présentée en exemple à la section 2.1. Pour le test de l'hypothèse H_0 contre H_1, cette statistique, notée $s_2(n)$, a pour expression :

$$s_2(n) = \frac{\sup_{\theta\in[0,1]} p_\theta(y_1,\cdots,y_n)}{p_0(y_1,\cdots,y_n)} \tag{2.10}$$

De la même manière que pour les précédents tests, le test correspondant à $s_2(n)$, est défini à l'aide de sa fonction de rejet notée $\varphi_2(n)$:

$$\varphi_2(n) = \mathbf{1}_{\{s_2(n)>\lambda_2(n)\}} \tag{2.11}$$

où le seuil $\lambda_2(n)$ peut être choisi en fonction d'un niveau α requis.

Le test du rapport de vraisemblance généralisé présente l'avantage notable, par rapport au test basé sur $s_1(n)$ (ou $\hat{\theta}_n$), d'être applicable indépendamment des contraintes sur la probabilité *a priori* de la distribution cible, sous l'hypothèse H_1.

2.2.3 Comportement asymptotique des statistiques de test

Pour étudier les propriétés asymptotiques des statistiques $s_1(n)$ et $s_2(n)$, on définit une suite $\{Y_n,n \geq 1\}$ de v.a. i.i.d., de loi admettant la densité p_θ, où $\theta \in [0,1]$. On note $S_1(n)$ l'estimateur du maximum de vraisemblance de θ :

$$S_1(n) = \underset{\alpha\in[0,1]}{\arg\max}\ p_\theta(Y_1,\cdots,Y_n) \tag{2.12}$$

et $S_2(n)$ la statistique du rapport de vraisemblance généralisé

$$S_2(n) = \frac{\sup\limits_{\theta\in[0,1]} p_\theta(Y_1,\cdots,Y_n)}{p_0(Y_1,\cdots,Y_n)} \tag{2.13}$$

L'information de Fisher, notée $I(\theta_0)$ en un point θ_0, joue un rôle important dans l'étude des propriétés asymptotiques des statistiques liées au maximum de vraisemblance. Elle est par définition

$$I(\theta_0) = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log p_\theta(Y_1)\right)^2_{|\theta=\theta_0}\right] \tag{2.14}$$

Comme sous la loi P_{θ_0}, le support de Y_1 ne dépend pas de θ_0

$$I(\theta_0) = \mathbb{E}_{\theta_0}\left[-\frac{\partial^2}{(\partial\theta)^2}\log p_\theta(Y_1)_{|\theta=\theta_0}\right] \tag{2.15}$$

Ici,

$$-\frac{\partial^2}{(\partial\theta)^2}\log p_\theta(Y_1) = \left[\frac{p_1(Y_i)-p_0(Y_i)}{p_\theta(Y_i)}\right]^2$$

et les distributions P_1 et P_0 sont différentes, donc $I(\theta_0)$ est strictement positif pour tout $\theta_0 \in [0,1]$.

Le théorème suivant donne, pour les diffŕentes valeurs du vrai paramètre de distribution θ, la loi asymptotique de $\sqrt{n}(S_1(n) - \theta)$ quand n tend vers $+\infty$.

Théorème 5 (Comportement de S_1)
On suppose que pour tout $\theta \in [0,1]$, l'Information de Fisher $I(\theta) < +\infty$, et que les deux espérances $\mathbb{E}_{P_0}\left[\left|\log \frac{p_1(Y_1)}{p_0(Y_1)}\right|\right]$ et $\mathbb{E}_{P_1}\left[\left|\log \frac{p_1(Y_1)}{p_0(Y_1)}\right|\right]$ sont finies. Alors :

– *sous P_θ, $\theta \in]0,1[$,*

$$\sqrt{n}\,(S_1(n) - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,I(\theta)^{-1}) \tag{2.16}$$

– *sous P_0,*

$$\sqrt{n}\,S_1(n) \xrightarrow{\mathcal{L}} \textit{le mélange} \begin{cases} \left|\mathcal{N}(0,I(0)^{-1})\right| & \textit{avec proba } \frac{1}{2} \\ 0 & \textit{avec proba } \frac{1}{2} \end{cases} \tag{2.17}$$

– *sous P_1,*

$$\sqrt{n}\,(S_1(n) - 1) \xrightarrow{\mathcal{L}} \textit{le mélange} \begin{cases} -\left|\mathcal{N}(0,I(1)^{-1})\right| & \textit{avec proba } \frac{1}{2} \\ 0 & \textit{avec proba } \frac{1}{2} \end{cases} \tag{2.18}$$

Théorème 6 (Comportement de S_2 sous H_0)
On suppose que l'Information de Fisher en $\theta = 0$ est finie, $I(0) < +\infty$, et que $\mathbb{E}_{P_0}\left[\left|\log \frac{p_1(Y_1)}{p_0(Y_1)}\right|\right] < +\infty$. Alors :

$$2\log S_2(n) \xrightarrow{\mathcal{L}} \textit{le mélange} \begin{cases} \chi_1^2 & \textit{avec proba } \frac{1}{2} \\ 0 & \textit{avec proba } \frac{1}{2} \end{cases} \tag{2.19}$$

2.2.4 Résultats d'optimalité de tests

Absence de test optimal à n fixé

Pour le test d'hypothèses simples, de $H_0(n) = \{\theta_n^\star = 0\}$ contre $H_1(n) = \{\theta_n^\star = \theta_1\}$, où θ_1 est fixé dans $]0,1]$, le test du rapport de vraisemblance est Uniformément Plus Puissant. Pour un seuil λ_n, la région de rejet de ce test est donnée par :

$$\left\{ \log \frac{p_{\theta_1}(y_1,\cdots,y_n)}{p_0(y_1,\cdots,y_n)} = \sum_{i=1}^{n} \log \frac{\theta_1 p_1(y) + (1-\theta_1)p_0(y_i)}{p_0(y_i)} > \lambda_n \right\}$$

On voit que ce test dépend de θ_1, donc à n fixé, il n'existe pas de test Uniformément Plus Puissant, de $H_0(n)$ contre $H_1(n)$. Les tests vont maintenant être évalués asymptotiquement.

Existence d'un test localement optimal

Le théorème 2 donne sous les hypothèses de régularité en annexe B.2, un test Localement Asymptotiquement Plus Puissant, pour tester "$\theta_n^\star = 0$" contre "$\theta_n^\star = \frac{\delta}{\sqrt{n}}$", où $\delta \in]0,1]$. Comme la fonction $\theta \longrightarrow p_\theta(y)$ est infiniment dérivable pour tout y, les conditions de régularités portent essentiellement sur les densités $p_1(y)$ et $p_0(y)$.

2.2.5 Conclusion

Nous avons présenté trois approches possibles pour effectuer la détection d'une classe de sons dans un document sonore et nous avons établi certaines propriétés statistiques des tests correspondants. Il apparaît que le Rapport de Vraisemblance Généralisé est le seul de ces 3 tests qui garantit l'invariance de la loi asymptotique sous l'hypothèse d'absence de la classe de sons, quelles que soient les distributions cible et non-cible.

Chapitre 3

Algorithmes de suivi d'une classe de sons dans un document

Le suivi d'une classe de sons consiste à indiquer les plages d'un document sonore, dans lesquels est présente la classe de sons en question. Beaucoup d'éléments définissant le problème de détection abordé au chapitre 2 se retrouvent également dans le suivi.

On se place alors dans le même contexte qu'au chapitre 2, où $\{y_1, \cdots, y_n\}$ est l'ensemble des vecteurs acoustiques extraits du document sonore, P_1 la distribution des vecteurs acoustiques, et P_0, celle correspondant à la classe non-cible. On suppose aussi que P_1 et P_0 admettent des densités, respectivement p_1 et p_0. Le suivi revient alors à (tenter de) déterminer la distribution dont est issue chaque observation y_t, pour $t = 1, \cdots, n$. C'est un problème de reconnaissance de la séquence des classes successives dont sont issues les observations.

Le résultat d'un système de suivi d'une classe de sons est une suite de segments définis par leurs dates de débuts et de fin, et un score associé à chaque segment. La décision de classifier un segment, comme provenant de la classe cible ou non, est prise en comparant la valeur du score à un seuil. Les performances d'un système de suivi seront visualisées à l'aide d'une courbe DET [Martin et al., 1997], qui représente la proportion d'observations non détectées comme appartenant à la classe cible en fonction de la proportion de fausses alarmes (observations détectées alors qu'elles ne font pas partie de la classe cible), ces chiffres évoluant en sens inverse l'un de l'autre, lorsque l'on fait varier le seuil de décision.

Nous allons d'abord décrire dans la section 3.1 l'approche générale utilisée en suivi. La section 3.2 présente des expériences en suivi de parole dans un document audiovisuel. D'autres expériences en suivi du locuteur dans une conversation téléphonique, menées dans le cadre de la campagne d'évaluations NIST (National Institute of Standards and Technologies), sont également présentées au chapitre 4.

3.1 Approche générale utilisée en suivi

Les différentes méthodes de suivi proposées par les auteurs peuvent se décomposer en deux phases. Elles consistent d'abord à détecter les transitions (ruptures) d'une classe de sons à une autre, qui se sont produites dans la séquence de vecteurs acoustiques extraits du document. Cela peut être effectué en utilisant, par exemple, un des algorithmes de détection de rupture sachant qu'il y a deux distributions des segments (présentés à la partie II). Ensuite, les segments ainsi trouvés (supposés homogènes) sont classifiés, en utilisant une des méthodes de détection d'une classe de sons (présentés au chapitre 2).

Deux types d'erreurs peuvent se produire durant la phase de détection de ruptures : la non-détection (omission) et la fausse alarme (insertion). En général, une omission est plus lourde de conséquences sur les performances de reconnaissance qu'une insertion, car la présence d'une rupture dans un segment conduit à une erreur systématique sur l'une ou l'autre des classes d'observations. Ainsi, certaines méthodes s'appuient sur une sur-segmentation du signal, voire l'utilisation de segments réguliers de très petite durée, bien qu'il y ait implicitement un fort taux de fausses alarmes de rupture (ce qui ne rentre pas en ligne de compte dans l'évaluation d'un système de suivi).

Dans ce contexte de forte présence de fausses alarmes de rupture, les frontières entre les segments de décision ne correspondent pas la plupart à de vraies ruptures. Il est alors opportun de calculer le score d'un segment sur tout un voisinage du segment, afin de préserver une certaine régularité en fonction des plages de décision.

Plusieurs stratégies de régularisation peuvent être utilisées. Par exemple, si la décision est prise sur des segments disjoints et réguliers de $(2w + 1)$ observations, le score peut être calculé sur de plus grandes fenêtres de lissage de taille $(2l + 1)$, avec $l > w$ (donc qui se recouvrent), centrées sur les mêmes instants que les segments. Dans le calcul du score, les observations peuvent également être pondérées, avec une pondération décroissante au fur et à mesure qu'on s'éloigne du centre de la fenêtre de lissage. Soit $h(i)$, $i = -l, \cdots, +l$, une fonction de pondération, l'expression du score du segment centré en t, basé sur le logarithme du rapport de vraisemblance entre la distribution cible P_0 et la distribution P_0, est donnée par :

$$\sum_{i=-l}^{+l} h(i) \ \log \ \frac{p_1(y_{t+i})}{p_0(y_{t+i})}$$

Des fonctions de pondération de Hamming ou triangulaire peuvent être utilisées.

Il est important de mentionner l'existence de certaines méthodes de suivi, qui proposent d'estimer directement au maximum de vraisemblance, selon un modèle de distribution, la séquence (ou chaîne) des classes des obvervations. Ainsi, certains auteurs comme [Gauvain et al., 1999, Meignier et al., 2000] proposent une modélisation de la distribution des vecteurs acoustiques du document par un Modèle de Markov Caché (MMC). Du reste, les MMC sont très utilisés en Traitement du Signal Sonore, notamment en Reconnaissance Automatique de la Parole, domaine dans lequel depuis plusieurs années, ils fournissent les meilleures performances.

3.2 Expériences en suivi de parole

Ces expériences portent sur le suivi de parole dans un document audiovisuel. La base de données utilisée est celle du corpus AIM, collectée par l'INA (Institut National de l'Audiovisuel).

3.2.1 La base de données AIM

Le corpus AIM est constitué de six bandes sonores :

- AIM1 : un documentaire de 35 minutes environ;
- AIM2 : une compilation de bandes annonces et de publicités d'environ 14 minutes;
- AIM3 : des journaux télévisés des chaînes M6 et Canal+, d'une durée de 12 minutes environ;
- AIM4 : un journal de la chaîne ARTE contenant des pages publicitaires, d'une durée de 13 minutes environ;
- AIM5 : un journal du soir de la chaîne France3 (Soir 3), d'une durée de 15 minutes environ;
- AIM6 : une sélection d'extraits de journaux de 20 heures de la chaîne TF1, d'une durée de 14 minutes environ.

Les classes de sons que l'on retrouve dans ces documents sont essentiellement la parole, la musique, la parole sur un fond de musique et une infime proportion d'autres sons (cris, rires, bruits de moteurs, coups de marteaux, bruits ambiants dûs aux conditions de prises de sons, etc...). Le tableau 3.1 donne les durées totales des différentes classes de sons, pour les six bandes sonores.

Bande son	Parole seule	Parole sur fond musical	Musique seule	Autres	Total
AIM1	18 mn 29 s	08 mn 05 s	05 mn 27 s	03 mn 34 s	35 mn 35 s
AIM2	01 mn 11 s	06 mn 11 s	05 mn 01 s	02 mn 20 s	14 mn 43 s
AIM3	09 mn 34 s	00 mn 40 s	01 mn 53 s	00 mn 00 s	12 mn 23 s
AIM4	08 mn 52 s	01 mn 53 s	01 mn 46 s	00 mn 46 s	13 mn 17 s
AIM5	12 mn 44 s	01 mn 37 s	01 mn 05 s	00 mn 00 s	15 mn 26 s
AIM6	11 mn 14 s	02 mn 23 s	01 mn 10 s	01 mn 58 s	14 mn 49 s

TAB. 3.1 – *Durées totales des différentes classes de sons, pour les six bandes sonores du corpus AIM.*

3.2.2 Suivi de parole dans une émission documentaire

Nous nous intéressons ici à la localisation des plages où est présente de la parole dans l'émission documentaire. Il s'agit de la segmentation du document en plages de parole éventuellement sur un fond musical, versus plages de musique seule. Précisons que la classe musique seule contient également de la chanson accompagnée d'instruments musicaux.

Données expérimentales

Les tests sont effectués sur la bande AIM1, qui est celle d'un documentaire d'environ 35 minutes, sans tenir compte des plages ne contenant ni de la parole ni de la musique (considérées comme du "bruit"), qui durent environ 3 mn 30 s. Les données d'apprentissage sont constituées des bandes AIM2, AIM3, AIM4, AIM5 et AIM6, soit 43 mn 35 s de parole seule, 12 mn 44 s de parole sur fond musical, et 10 mn 55 s de musique seule (ce qui est fort peu, compte tenu de la diversité du signal musical).

Paramétrisation acoustique

Les bandes sonores sont échantillonnées à la fréquence de 22050 Hz. Pour chaque trame d'une durée de 10 ms, 16 coefficients cepstraux et leurs deltas sont extraits, soit un vecteur acoustique de 32 paramètres. Notons que le delta du log de l'énergie pourrait être inclus dans la paramétrisation. Le rajout de ce dernier n'affecte pas de manière significative les performances (pour des modèles Gaussiens). La soustraction de la moyenne des paramètres à l'échelle de l'enregistrement, connue comme une technique classique et efficace de compensation de canal de transmission, n'est pas appliquée, compte tenu de la variabilité du canal sur certains enregistrements, notamment sur les bandes qui sont des compilations de différentes émissions. Pour ce type d'enregistrement, des méthodes de compensation de canal, comme la soustraction de la moyenne sur une fenêtre glissante où le filtrage RASTA, sont plus adaptées.

Modélisation

Pour chacun de trois types de sons, parole seule, parole sur fond musical et musique seule, un MMG (avec des matrices de covariance pleines) est appris. Le modèle de présence de parole est construit par mélange équiproprable du modèle de parole seule et du modèle de parole sur fond musical. Des MMG comprenant 1, 5 et 10 Gaussiennes sont testés.

Score et décison

Chaque 10 ms (durée de la trame), une décision est prise en comparant le score à un seuil. Les trois scores présentés à la section 2 sont testés pour une durée de fenêtre de lissage égale à 10 cs, 25 cs, 50 cs, 1 s, 2 s, 5 s et 10 s.

Résultats et discussion

Les figures 3.1, 3.2 et 3.3 représentent les performances obtenues avec respectivement les modèles d'ordre 1, 5 et 10. Dans chacune de ces figures, les performances sont représentées pour les trois scores et pour les différentes durées de la fenêtre de lissage.

En comparant les trois figures, on constate que les modèles Gaussiens fournissent des performances légèrement meilleures, surtout pour les courtes durées de fenêtre de lissage. Cela pourrait s'expliquer par une sur-spécialisation des MMG d'ordre supérieur à 1 aux données d'apprentissage. Les commentaires qui vont suivre portent donc sur les résultats obtenus avec les modèles Gaussiens.

Indépendamment du score, les taux d'erreurs ont un comportement généralement prévisible en fonction de la durée de la fenêtre de lissage. Ainsi, quand on augmente la durée de la fenêtre, les taux d'erreurs décroissent d'abord de manière monotone jusqu'à un certain seuil, puis elles s'accroissent.

Pour les courtes durées de fenêtre, le rapport de vraisemblance donne de moins bonnes performances. Quand la durée de la fenêtre est comprise entre 1 s et 5 s, les performances des trois scores sont sensiblement comparables. Pour les trois scores, la durée optimale de la fenêtre est entre 2 s et 5 s, avec un EER autour de 12%.

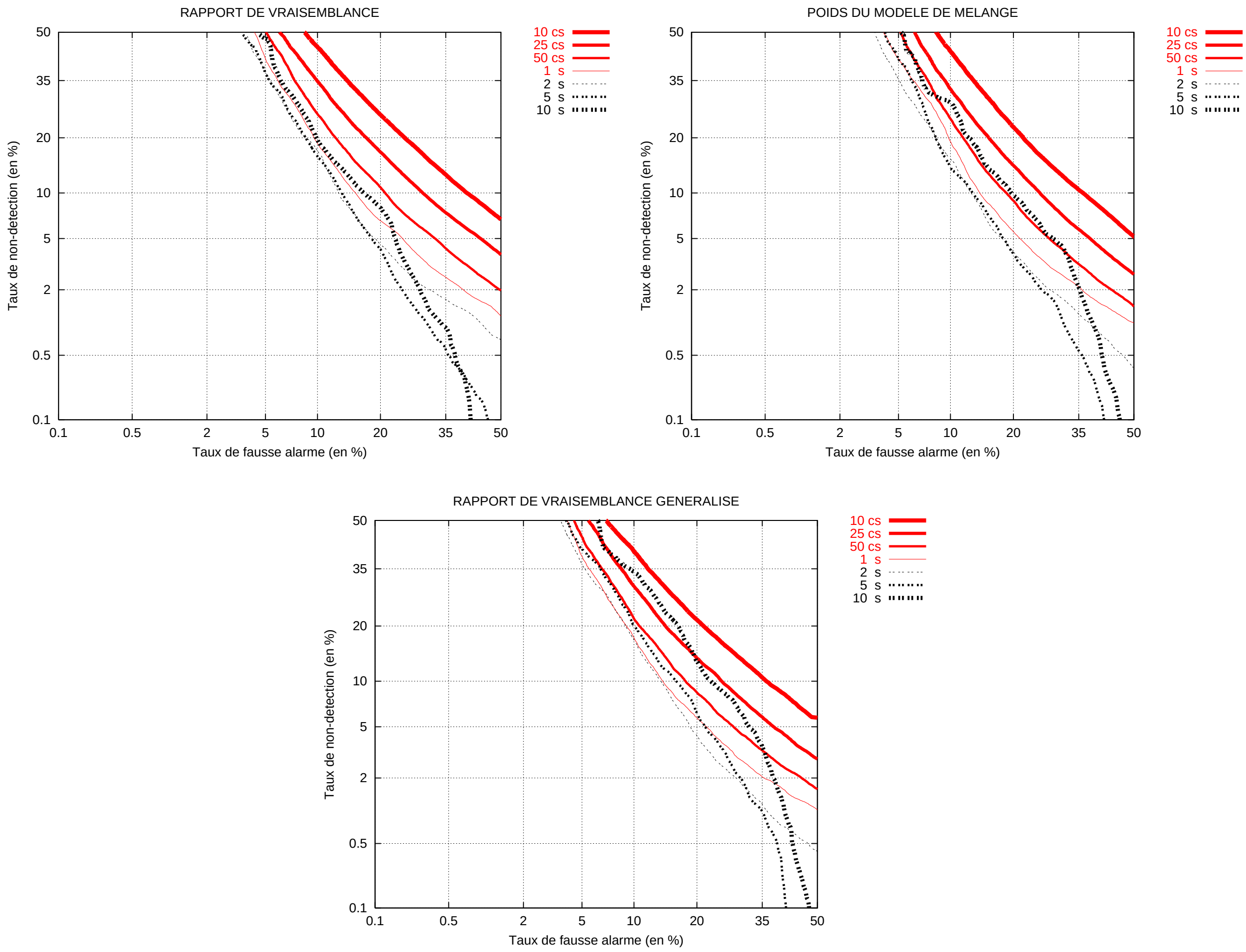

FIG. 3.1 – *Performances des trois scores en suivi de parole dans une émission documentaire pour différentes durées de fenêtres de lissage. Les modèles de parole seule, de parole sur fond musical et de musique seule, sont Gaussiens.*

La figure 3.4 représente, pour le poids du modèle de mélange lorsque l'on utilise une fenêtre de 2 s, les fonctions de répartition de la parole seule, de la parole sur fond musical et de la présence de parole, et le complémentaire à 1 de la fonction de répartition de la musique seule.

La figure 3.5 donne sur le diagramme supérieur la structure segmentale du documentaire, selon les trois types de sons.Les zones noires correspondent aux plages de parole seule, les zones grises aux plages de parole sur fond musical, et les zones blanches aux plages de musique

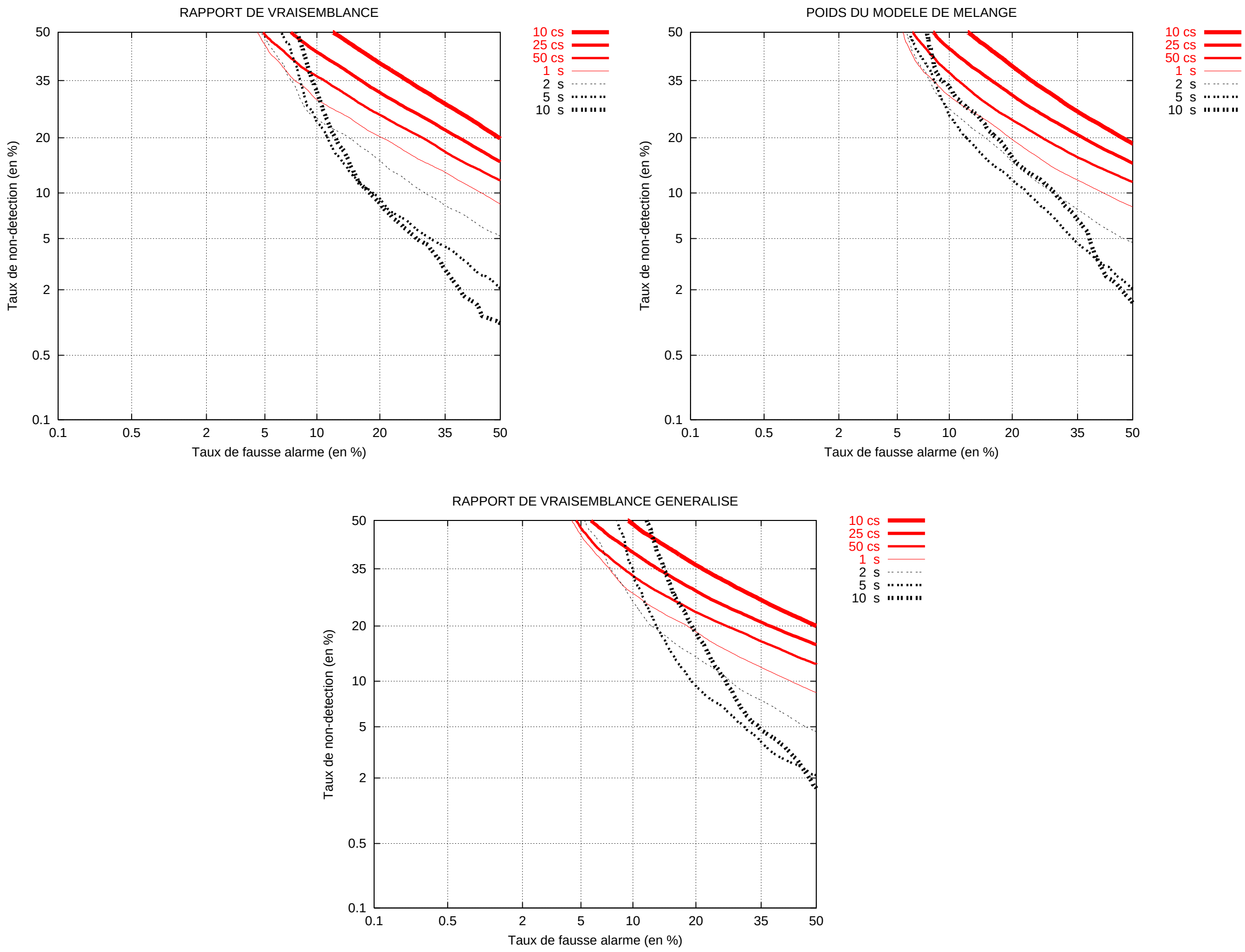

FIG. 3.2 – *Performances des trois scores en suivi de parole dans une émission documentaire pour différentes durées de fenêtres de lissage. Les modèles de parole seule, de parole sur fond musical et de musique seule, sont des MMG d'ordre* 5.

seule. Le diagramme inférieur correspond à la segmentation obtenue avec le score du modèle de mélange, en utilisant une fenêtre d'analyse de 2 s et en prenant une décision binaire à l'EER (qui est de 12%). Sur ce diagramme, les zones noires représentent les plages où est présente de la parole (éventuellement sur fond musical), et les zones blanches les plages de musique seule.

On constate sur la figure 3.4, comme on pouvait s'y attendre, que la distribution des scores

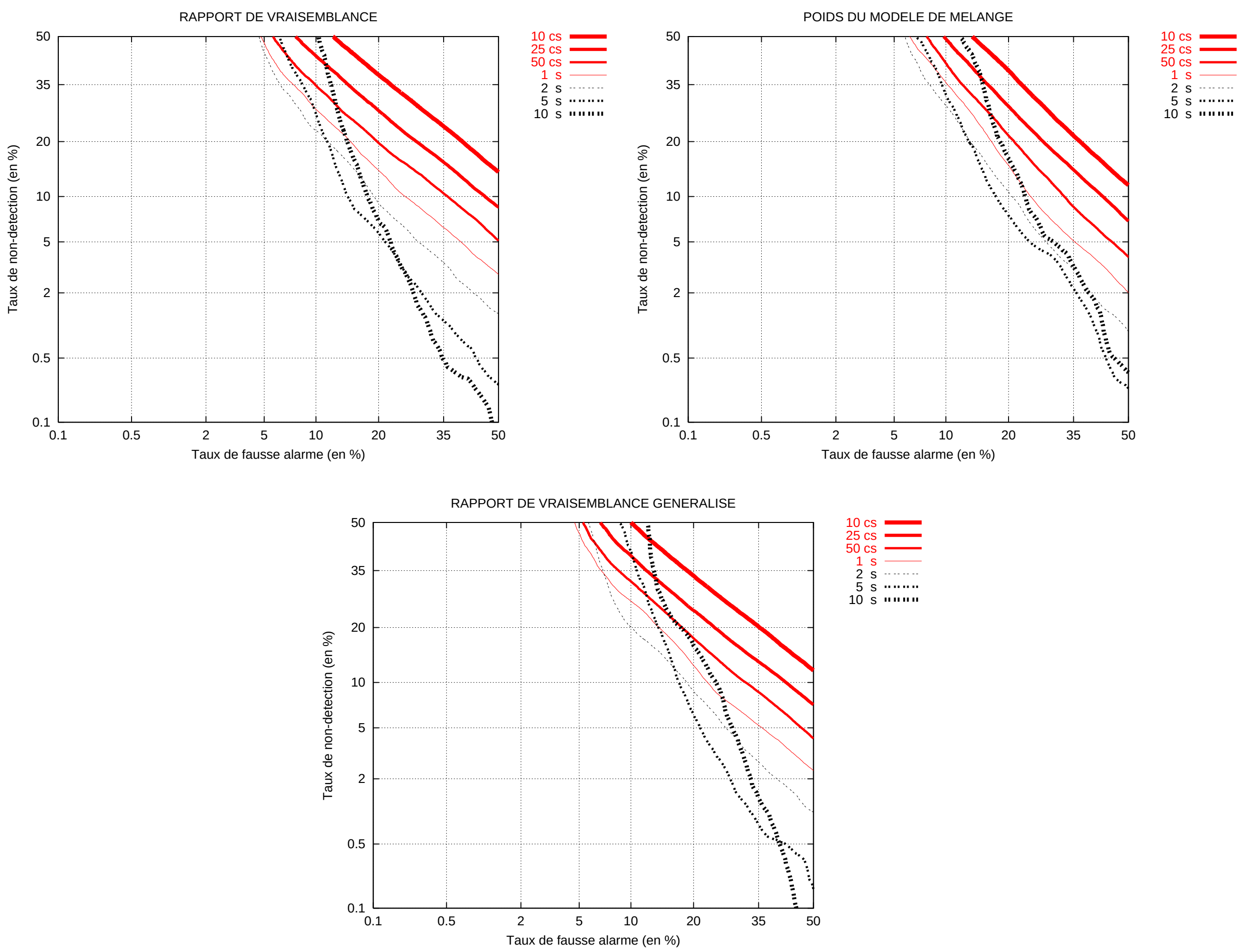

FIG. 3.3 – *Performances des trois scores en suivi de parole dans une émission documentaire pour différentes durées de fenêtres de lissage. Les modèles de parole seule, de parole sur fond musical et de musique seule sont des MMG d'ordre* 10.

de parole sur fond musical est plus proche que celle des scores de parole seule, de la distribution des scores de musique seule. Les erreurs sont donc plus dues aux plages de parole sur fond musical, qu'a celles de parole seule. A ce propos, on peut voir sur la figure 3.5, les erreurs de reconnaissance commises sur les trois dernières plages de parole sur fond musical (en gris).

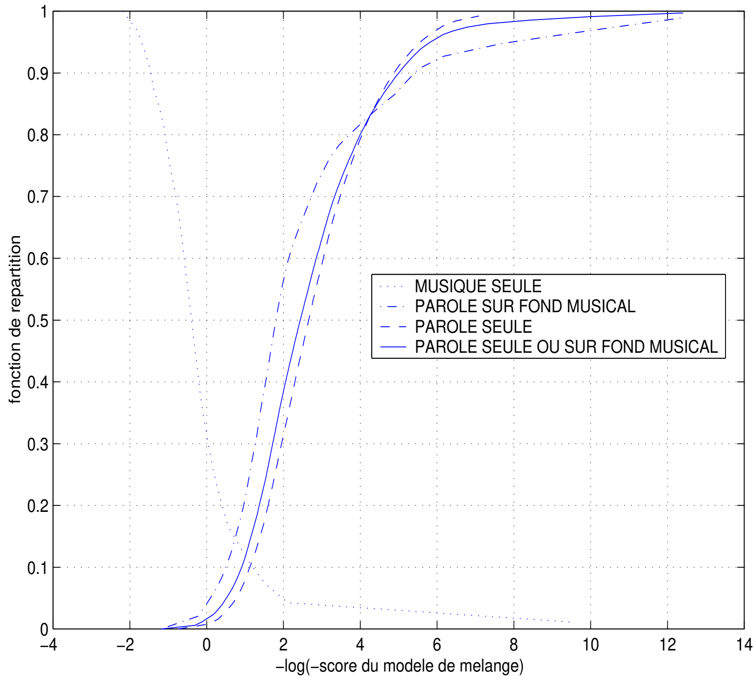

FIG. 3.4 – *Fonctions de répartion du poids du modèle de mélange, pour les plages de parole seule, parole sur fond musical et présence de parole, et le complémentaire à 1 de la fonction de répartition de ce même score, pour les plages de musique seule*

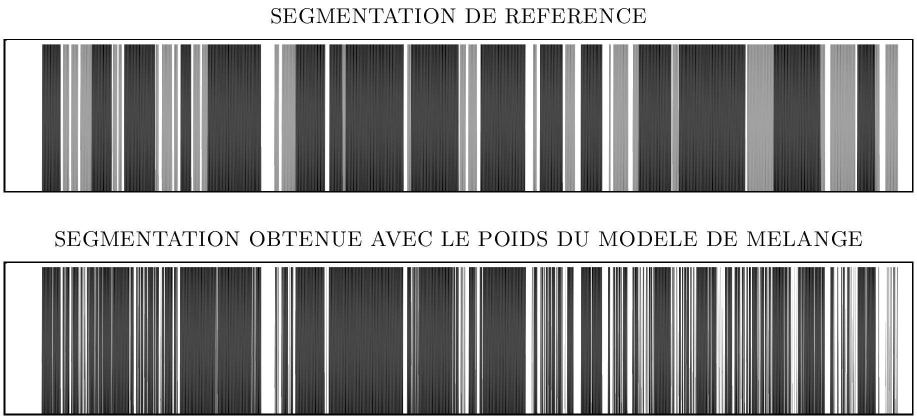

FIG. 3.5 – *Visualisation de la segmentation obtenue avec le poids du modèle de mélange. Sur le diagramme supérieur, la segmentation de référence de la bande AIM1 (une émission documentaire) : en blanc les plages de musique seule, en gris les plages de parole sur fond musical et en noir les plages de parole seule. Sur le diagramme inférieur, la segmentation obtenue en utilisant le score du modèle de mélange, avec une fenêtre d'analyse de 2 s et en prenant une décision binaire à l'EER (qui est ici de 12 %).*

3.3 Conclusion

Pour le problème traité dans ce chapitre, c'est à dire la détection parole/musique, les modèles MMG (d'ordre supérieurs à 1) n'améliore pas les performances par arpport aux modèles Gaussiens. Il convient toutefois de rappeler qu'il s'agit de modèles à covariances pleines, ce qui a pour conséquence de faire croitre rapidement le nombre de paramètre du modèle avec son ordre. Par ailleurs on observe que les trois scores testés fournissent des performances à peu près comparables, au voisinage de 10% d'EER.

Chapitre 4

Détection et suivi du locuteur

L'objet de ce chapitre est d'évaluer et de comparer sur des tâches de détection et de suivi du locuteur les méthodes de détection et de suivi d'une classe de sons présentées aux chapitres 2 et 3. La classe de sons cible est ici la parole d'un locuteur fixé, qui est appelé *locuteur cible*. La vérification du locuteur consiste à dire si un locuteur cible donné est présent dans un document sonore. Quant au suivi, il consiste à localiser les plages sonores où le locuteur cible est présent dans le document.

Ces travaux ont été effectués dans le cadre des évaluations annuelles organisées par l'institut américain NIST (National Institute of Standards and Technologies) et qui portent sur la Reconnaissance Automatique du Locuteur (RAL). L'équipe Sigma2 de l'IRISA participe aux évaluations NIST [NIST, site] en RAL en 1998, 1999 et 2000, en tant que membre du consortium ELISA [ELISA, 2000].

Les trois principales tâches des évaluations NIST en RAL, et auxquelles participe l'IRISA sont :

- la vérification d'un locuteur cible dans un énoncé mono-locuteur,
- la vérification d'un locuteur cible dans une conversation entre deux individus,
- et le suivi du locuteur dans une conversation entre deux individus.

Ces tâches sont respectivement connues sous les appellations NIST, *one-speaker detection*, *two-speaker detection* et *speaker tracking*. Dans ce manuscrit, elles seront alors désignées par les appelations *vérification du locuteur*, *détection du locuteur* et *suivi du locuteur*.

En 1998, la participation de l'IRISA aux évaluations NIST était restreinte à la vérification du locuteur. Depuis 1999, les tâches auxquelles participe l'IRISA se sont élargies à la détection et au suivi du locuteur. Les travaux qui sont exposés ici portent essentiellement sur la détection du locuteur et le suivi du locuteur durant les évaluations des années 1999 et 2000.

4.1 Méthodologie générale pour la détection et le suivi du locuteur

La méthodologie générale utilisée en détection et en suivi du locuteur s'inscrit bien dans le schéma des systèmes présentés au chapitre 1 de la partie I. On rappelle que ce schéma propose une décomposition d'un système en quatres modules :

- un module de paramétrisation acoustique,
- un module de modélisation,
- un module de calcul des scores,
- et un module de décision.

Les sections suivantes présentent ces différents modules dans le cas précis de la détection et du suivi du locuteur.

4.1.1 Paramétrisation acoustique

Les signaux de parole sont paramétrisés et sont représentés par un ensemble de vecteurs de paramètres acoustiques. Les paramètres acoustiques utilisés sont des coefficients cepstraux, basés sur des coefficients de bancs de filtres spectraux triangulaires, disposés selon une échelle Mel, présentés à la section 1.1 de la partie I.

4.1.2 Modélisation

Pour chaque locuteur cible X, un modèle de distribution des vecteurs acoustiques (également noté X) est estimé à partir de données d'apprentissage. Dans ces expériences les données d'apprentissage sont d'une durée de 2 minutes par locuteur. En complément du modèle du locuteur, un modèle des vecteurs acoustiques *non* produits par le locuteur est également appris. Pour éviter d'apprendre ce dernier modèle pour chaque locuteur cible, dans la pratique, on estime un seul modèle Ω indépendant du locuteur (approche dite du *modèle du monde*) à partir d'une grande quantité de données (environ 2 heures) ne contenant de la parole d'aucun des locuteurs cibles (ni des imposteurs) testés.

Selon les notations utilisées aux chapitres 2 et 3, un modèle du locuteur correspond au *modèle cible* P_1, et le modèle du monde Ω joue le rôle du *modèle non-cible* P_0. Pour des raisons d'ordre pratique, le modèle du monde est construit par mélange équiprobable de 4 sous-modèles du monde dépendant du type de combiné et du genre des individus. Les sous-modèles du monde sont utilisés individuellement comme modèle non-cible en vérification du locuteur, où l'on suppose connu le type de combiné téléphonique utilisé et le genre de l'individu que l'on cherche à détecter sur le document de test.

La famille de modèles utilisée est, comme dans le chapitre précédent, celle des Modèles de Mélanges de Gaussiennes (MMG). Toutefois, les Gaussiennes sont ici contraintes à avoir des matrices de covariance diagonales. Les modèles sont estimés en utilisant l'algorithme EM, avec 10 itérations. Ce choix correspond à l'état de l'art actuel dans le domaine de la RAL, comme le reflète le numéro spécial du journal *Digital Signal Processing* [DSPJ, 2000], consacré aux évaluations NIST 1999.

4.1.3 Calcul du score et décision

Dans une phase de test, un score pour le document à tester est calculé en utilisant le modèle du locuteur cible X et le modèle non cible Ω. Le calcul du score est un module sur lequel les systèmes de détection et de suivi diffèrent. Les scores sont calculés en utilisant les méthodes de détection et de suivi décrites au chapitre 2 pour la détection et au chapitre 3 pour le suivi. Dans ces chapitres, $\{y_1, \cdots, y_n\}$ correspond aux n vecteurs de paramètres acoustiques extraits du document de test, et p_1 et p_0 représentent respectivement les densités des distributions cible et non-cible X et Ω.

La décison de présence ou non du locuteur cible dans le document est prise par comparaison de la statistique de test à un seuil λ. Dans ces expériences, ce seuil est indépendant du document de test et du locuteur cible auquel il est confronté.

4.2 Bases de données 1999 et 2000

Les données utilisées dans ces expériences sont celles des évaluations NIST et sont extraits de la base Switchboard-2, collectées par le *Linguistic Data Consortium* (LDC). De plus amples informations sur le LDC peuvent être recueillies sur son site web [LDC, site]. Ces données sont en grande partie constituées d'enregistrements téléphoniques d'étudiants du sud des États Unis. Il existe deux types de combinés téléphoniques aux États : le type électret et le type carbone. Les enregistrements sont à une large majorité de type électret.

Pour chaque locuteur cible, les données d'apprentissage sont constituées de 2 minutes de parole téléphonique provenant du même numéro d'appel. Une différence subsiste entre les données d'apprentissage des évaluations 1999 et 2000. Pour les évaluations de 1999, les 2 minutes de parole ont été obtenues à partir de deux sessions d'enregistrement d'une minute chacune, tandis qu'en 2000, elles sont extraites d'une seule et même session d'enregistrement.

Les documents de test de la vérification du locuteur sont constitués d'enregistrements de l'une des deux parties d'une conversation, et sont d'une durée variant de quelques (environ 3) secondes à une minute. Chaque document est testé contre 11 locuteurs cibles, dont le locuteur présent dans le document, et sont tous du même genre que ce dernier. Lors du test d'une des 11 identités, il est interdit d'utiliser le résultat obtenu avec les 10 autres identités cibles, ce qui correspond bien à de la vérification du locuteur (acceptation/rejet pour chacune des 11 identités, indépendamment) et non à de l'identification en ensemble ouvert (choisir 1 parmi 11 ou aucun).

Les documents de test pour la détection du locuteur sont obtenus en réunissant les deux parties d'une conversation entre deux individus et sont chacun d'une durée totale d'une minute (± 1 seconde). La proportion de présence du locuteur cible dans le document de test varie de 0 à 100%. Chaque document est testé contre 22 locuteurs cibles, dont les deux locuteurs présents dans le document. Ces 22 locuteurs cibles sont constitués de deux sous-ensembles de 11 locuteurs, dont chacun est du même genre qu'un des locuteurs du document de test.

Le protocole d'évaluation du suivi du locuteur est un sous-ensemble de celui de la détection du locuteur. Chaque document de test est confronté à 4 locuteurs cibles (choisis parmi les 22 de la vérification du locuteur), dont généralement [1] les deux locuteurs cibles présents dans le document. Les genres des 4 locuteurs cibles sont répartis équitablement en deux sous-ensembles, selon les genres des locuteurs présents dans le document.

Le tableau 4.1 donne, pour les années 1999 et 2000, le nombre de documents de test en détection et en suivi du locuteur.

	Détection du locuteur	Suivi du locuteur
1999	1723	1000
2000	2748	1000

TAB. 4.1 – *Nombre de documents de test des évaluations NIST, en détection du locuteur et en suivi du locuteur, pour les années* 1999 *et* 2000.

1. Si toutefois les deux locuteurs présents dans le document de test ont des données d'apprentissage, afin de pouvoir être choisi comme locuteur cible.

4.3 Mesure de performances des systèmes

Les performances d'un système de détection et de suivi sont mesurées à l'aide d'une courbe DET (Detection Error Trade off), qui représente sur une échelle particulière, le taux d'erreur de non-détection en fonction du taux de fausse alarme, sur l'ensemble des documents de test et des locuteurs cibles auxquels les documents sont confrontés. Les erreurs sont comptabilisées par document de test en détection du locuteur, et par centième de seconde du document en suivi de locuteur. L'échelle de représentation des erreurs est telle que la courbe DET théorique correspondant à un score Gaussien (aussi bien sous la distribution cible que la distribution non-cible) est une droite sur le plan des deux axes d'erreurs [Martin et al., 1997]. Les scores que l'on calcule, sont en général[2] asymptotiquement Gaussiens, et si l'on dispose d'un nombre assez important de configurations de test, les courbes DET ont une allure à peu près rectiligne.

4.4 Les systèmes IRISA/ELISA aux évaluations NIST 1999

Cette section présente les systèmes IRISA/ELISA réalisés dans le cadre des évaluations NIST 1999 et 2000, pour les tâches de détection et de suivi du locuteur, et leurs résultats commentés [Seck et al., 2000]. La description des systèmes s'appuie sur la méthodologie générale présentée à la section 4.1, en spécifiant les grandeurs numériques particulières utilisées.

4.4.1 Spécification des paramètres acoustiques

Les vecteurs de paramètres acoustiques sont obtenus de la manière suivante. Une préaccentuation avec un facteur de 0,95 est appliquée au signal de parole. Ensuite, toutes les 10 ms, 16 coefficients cepstraux sont extraits sur une trame de 32 ms, à laquelle une fenêtre de Hamming est au préalable appliquée. Les coefficients spectraux sont des coefficients d'un banc de filtres triangulaires disposés selon une échelle Mel. Le vecteur de paramètres acoustiques de la trame est constitué des 16 coefficients cepstraux et de leurs deltas ; soit un vecteur de

2. Ces scores sont pour la plupart des basés sur une somme de variables aléatoires ou des estimateurs du maximum de vraisemblances. Sous l'hypothèse d'ind! épendance des variables, ces scores sont asymptotiquement Gaussiens.

32 paramètres.

Une soustraction de la moyenne des paramètres à l'echelle de chaque enregistrement est ensuite appliquée aux vecteurs de paramètres (soustraction cepstral à long terme). Il est important de mentionner que la soustraction cepstrale n'est pas théoriquement appropriée au document de test où sont simultanément présents deux canaux de transmission. La moyenne sur le document correspond à une moyenne pondérée des moyennes sur les deux canaux. Une soustraction cepstrale sur une fenêtre glissante est théoriquement plus appropriée.

4.4.2 Spécification des modèles de distribution

Les modèles de locuteurs aussi bien que les sous-modèles du monde sont des mélanges de 128 Gaussiennes à matrices de covariances diagonales, estimés avec 10 itérations de l'algorithme EM légèrement modifié (en utilisant un facteur de régularisation de 0,1 - voir chapitre sur l'estimation des MMG). Le modèle du monde, mélange équiprobable des quatre sous-modèles du monde (selon le type de combiné et le genre), est alors un mélange de 512 Gaussiennes à covariances diagonales.

4.4.3 Détection du locuteur

Pour la détection du locuteur, deux systèmes ont été présentés : un système officiel soumis dans les délais et un système secondaire qui a été mis en œuvre ultérieurement.

Système officiel

Le système officiel est basé sur un système classique de vérification du locuteur. Pour un locuteur X donné, ce système procède par un test du rapport de vraisemblance entre les distributions X et Ω. La fonction de rejet du test est $\varphi_0(n)$ (formule 2.7), avec la constante $\lambda_0(n) = \lambda_0^n$, où λ_0 est un réel positif.

Les taux d'erreur sont calculés sur l'ensemble des documents de test et des locuteurs cibles de la base de données d'évaluation, en faisant varier la constante λ_0 indépendament du locuteur et du document. Pour un locuteur X donné, le niveau (taux de fausse alarme) asymptotique théorique du test, noté $\alpha_0(X)$, a pour expression :

$$\alpha_0(X) = \begin{cases} 1 & \text{si} \quad \log(\lambda_0) < -K(\Omega,X) \\ \frac{1}{2} & \text{si} \quad \log(\lambda_0) = -K(\Omega,X) \\ 0 & \text{si} \quad \log(\lambda_0) > -K(\Omega,X) \end{cases} \tag{4.1}$$

où $K(\Omega,X)$ est l'information de Kullback de Ω par rapport à X. Il apparaît que si les informations de Kullback des distributions des locuteurs cibles par rapport à Ω ne sont pas les mêmes, ce qui est très vraisemblable, il existe des seuils pour lesquels le test n'a pas le même niveau asymptotique pour tous les locuteurs. Pour avoir un même niveau du test, indépendamment du locuteur cible, le score de ce système doit être normalisé.

Système secondaire

Le système secondaire, développé plus tard, n'a pas été soumis officiellement aux évalutions NIST 1999. Ce système est basé sur une modélisation de la distribution des blocs de vecteurs acoustiques du document de test, par un mélange de deux distributions : la distribution cible X et la distribution non-cible Ω. Le test du système est, sur le principe, celui défini par la fonction de rejet $\varphi_1(n)$ (formule 2.9), avec la constante $\lambda_1(n) = \lambda_1$, où λ_1 est un réel strictement positif. Sur le principe car, à la place du vecteur acoustique, le bloc de vecteurs acoustiques consécutifs joue le rôle d'observation.

Ainsi, pour un entier w et un instant t fixés, on dénote par $y_{\underline{w},t} = \{y_{t-w}, \cdots ,y_{t+w}\}$ le bloc des $(2w+1)$ vecteurs acoustiques consécutifs, centré en t. En supposant l'indépendance[3] des vecteurs acoustiques, la densité d'un bloc est le produit des densités des vecteurs acoustiques qui le composent. A cet effet, les densités du bloc, pour le modèle du monde et pour le modèle du locuteur cible, sont respectivement définies par :

$$p(y_{\underline{w},t}|\Omega) = \prod_{i=t-w}^{t+w} p(y_i|\Omega) \qquad \text{et} \qquad p(y_{\underline{w},t}|X) = \prod_{i=t-w}^{t+w} p(y_i|X)$$

3. Les vecteurs acoustiques ne sont pas indépendants, notammment quand ils sont proches, d'autant plus qu'il y a un chevauchement entre deux trames (morceaux du signal de parole sur lesquels sont extraits les vecteurs de paramètres acoustiques).

où $p(.|\mathcal{D})$ est la densité de la distribution $\mathcal{D}$. Une fois les densités des observations-blocs fixées, le modèle de mélange, l'estimateur du poids modèle cible dans ce mélange, ainsi que le test qui s'en suit, se définissent de la même manière que pour les observations-trames.

L'estimation du poids du modèle cible est effectuée sur tous les blocs du document de test, i.e. sur l'ensemble $\{y_{\underline{w},t}, t = w+1, \cdots, n-w\}$. On garde tout de même à l'esprit que l'hypothèse d'indépendance des observations(-blocs), supposée par l'algorithme EM pour estimer le poids du modèle cible, est bafouée, compte tenu du chevauchement très important entre les blocs. Il a semblé préférable d'avoir un plus grand nombre d'observations-blocs, au détriment de leur indépendance.
Plusieurs durées de blocs sont utilisées dans ces expériences.

Pour un seuil λ_1 strictement positif, le niveau asymptotique du test pour un locuteur X fixé, est égal à zéro (sous l'hypothèse d'indépendance des blocs), compte tenu de la loi asymptotique de $\sqrt{n}S_1(n)$, donnée par la formule 2.17. Le niveau est alors indépendant du locuteur.

Résultats et interprétations

Les résultats présentés, portent sur l'ensemble des données d'évaluations de la campagne NIST 1999 (section 4.2) pour la tâche de détection du locuteur.

Sur la figure 4.1, les différentes courbes DET représentent les performances du système secondaire, pour différentes durées de blocs exprimées en secondes. On observe un accroissement général des performances en fonction de la durée des blocs, mais les performances semblent se stabiliser avec des blocs d'une durée comprise entre 1 et 2 secondes. Cette observation peut s'expliquer par le fait que l'utilisation des blocs renforce la cohérence du modèle des vecteurs acoustiques du document, en apportant une information supplémentaire sur la longueur minimale des plages homogènes.

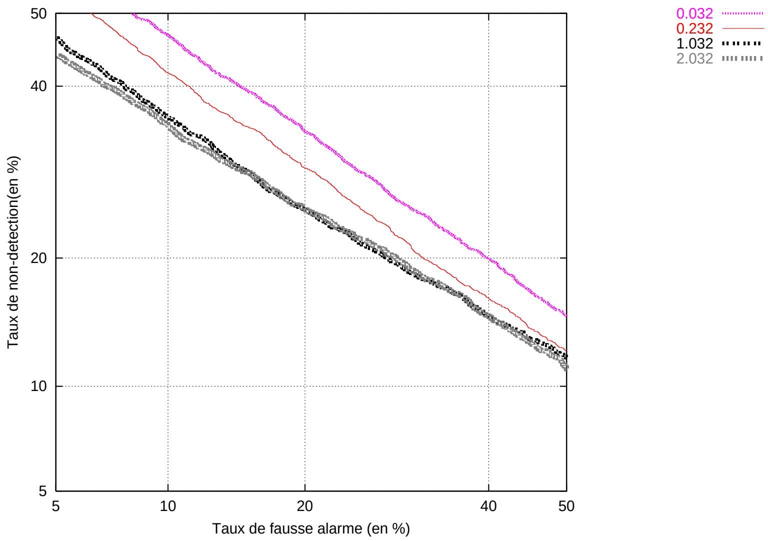

FIG. 4.1 – *Détection du locuteur. Performances du système secondaire pour differentes durées de blocs (exprimées en secondes)*

La figure 4.2 représente les performances obtenues avec le système officiel et celles du système secondaire pour deux durées des blocs : quand le bloc n'est constitué que d'une trame et quand le bloc a une durée approximative de 2 secondes (2.032 s). Le système secondaire donne de meilleures performances même quand le bloc n'est constitué que d'une seule trame.

La figure 4.3 compare les performances du système secondaire de l'IRISA avec celles des systèmes soumis par les autres laboratoires participant aux évaluations NIST 1999. Les systèmes des laboratoires 1 et 2 donnent les meilleures performances, ceux des laboratoires 3, 4 et le système secondaire de l'IRISA, présentent des niveaux de performances comparables, avec de légèrement moins bonnes performances du laboratoire 4.

Les différences entre le système secondaire de l'IRISA et les meilleurs systèmes présentés aux évaluations NIST portent essentiellement sur l'apprentissage du modèle du locuteur et sur le calcul du score. En effet, pour les autres systèmes, le modèle du locuteur est appris par adaptation à partir du modèle du monde et non par modélisation au maximum de

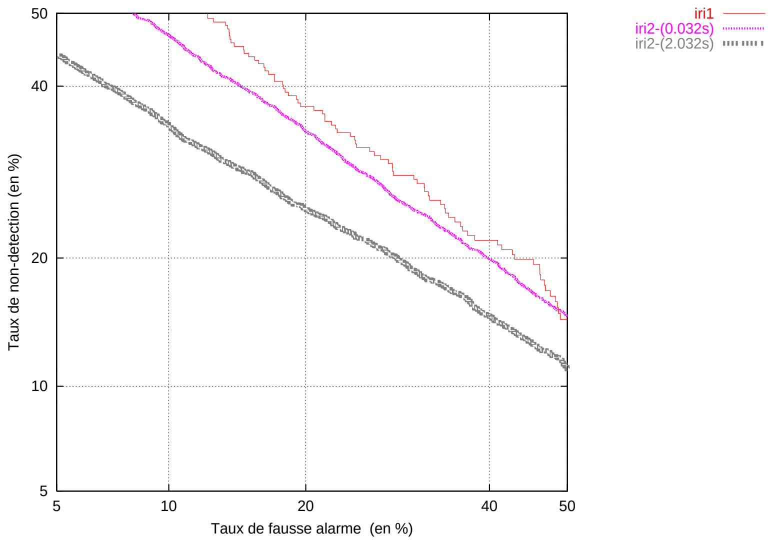

FIG. 4.2 – *Détection du locuteur. Performances du système officiel (iri1) et du système secondaire (iri2) quand le bloc correspond à une trame (*0.032 *seconde) et quand le bloc à une durée d'environ* 2 *secondes (*2.032 *s).*

vraisemblance des données d'apprentissage seulement.

En ce qui concerne le calcul du score, les meilleures systèmes calculent un logarithme du rapport de vraisemblance moyen entre le modèle du locuteur cible et le modèle du monde, sur un pourcentage fixé des meilleurs vecteurs acoustiques (au sens du plus grand rapport de vraisemblance). Ils utilisent également une normalisation des scores par rapport au type de combiné du vecteur acoustique de test qui, d'après leurs expériences de développement, est un facteur d'amélioration des performances de leurs systèmes.

La figure 4.4 représente les courbes DET du système secondaire avec une durée de bloc d'environ 2 secondes, pour les quatre conditions notées MF, FM, FF et MM, selon les genres des deux locuteurs présents dans le document de test, la première des deux lettres indiquant aussi le genre du locuteur cible. Par exemple, la notation MF signifie que le locuteur cible est du genre masculin et le document de test est une conversation entre un homme et une femme.

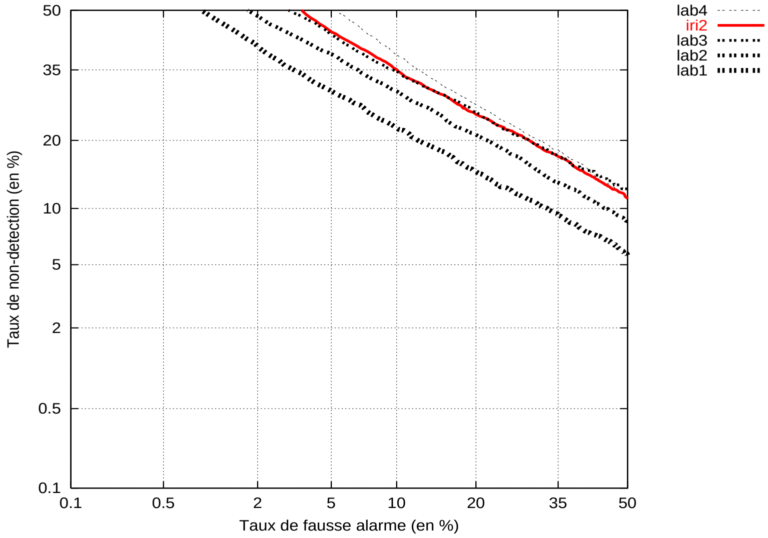

FIG. 4.3 – *Détection du locuteur. Performances du système secondaire quand le bloc est d'une durée d'environ 2 secondes (2.032 s), et celles des autres laboratoires participants.*

Pour les faibles taux de fausse alarme, les meilleurs taux de non-détection sont obtenus quand le locuteur cible est un homme, aussi bien dans le cas où les deux locuteurs du document de test sont du même genre (MM > FF), que quand ils sont de genres différents (MF > FM). En outre, notons presque avec surprise que les performances dans le cas MM sont meilleures que celles dans le cas FM, seulement sur le quart de plan inférieur droit de la figure, sans qu'il ne soit! possible, pour le moment, d'apporter une explication précise à cette observation.

4.4.4 Suivi du locuteur

Pour la tâche de suivi du locuteur, trois systèmes ont été réalisés dans le cadre des évaluations NIST 1999. Seul le second a été effectivement soumis. Le troisième système a été développé ultérieurement. Les systèmes vont être décrits en faisant référence à la méthodologie générale présentée à la section 3.1.

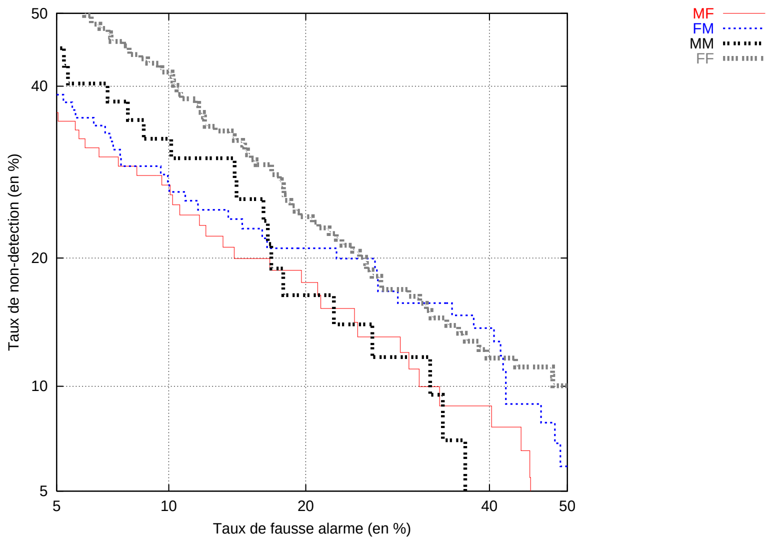

FIG. 4.4 – *Détection du locuteur. Performances du système secondaire quand le bloc est d'une durée d'environ 2 secondes (2.032 s), pour les quatre conditions dépendant des genres des deux locuteurs du document de test.*

Premier système de suivi du locuteur

Le premier système de suivi du locuteur utilise un calcul des scores et une décision sur des segments de durée fixe et constante. Ce système est basé sur un score calculé à chaque instant t comme la moyenne, pondérée par une fenêtre de Hamming, du log-rapport de vraisemblance sur une fenêtre de taille $2w+1$ centrée sur l'instant t.

$$\sum_{i=-w}^{w} h(i) \ \log \ \frac{p(y_{t+i}|X)}{p(y_{t+i}|\Omega)}$$

où $\{h(i), i = -w, \cdots, +w\}$ sont les coefficients d'une fenêtre de Hamming de longueur $2w+1$, normalisés de manière que leur somme soit égale à 1. Dans ces expériences, une longueur de fenêtre $(2w+1)$ correspondant à une durée d'environ 2 secondes, est utilisée. Afin de ne pas dépasser la taille maximale des fichiers de soumission fixée par NIST, les scores du système sont calculés sur des segments constitués de 5 trames (50 ms), en effectuant la moyenne sur les scores trame à trame. La décison finale est prise par comparaison du score du segment à un seuil.

Second système de suivi du locuteur

Le second système de suivi du locuteur effectue d'abord une segmentation en plages homogènes, en utilisant la méthode de détection de rupture présentée à la section 2.2.1 de la partie II. À chaque segment ainsi obtenu, est associé son score, qui est le logarithme du log-rapport de vraisemblance moyen des trames du segment. La moyenne est calculée en utilisant une fenêtre de Hamming normalisée, de longueur égale à celle du segment.

Troisième système de suivi du locuteur

Le troisième système de suivi du locuteur procède par un calcul de score et une prise de décision, sur des segments réguliers, et sur le même principe que le premier système. La seule différence entre le premier et le troisième système, est le score trame à trame utilisé. Le score trame à trame du troisième système utilise le score du second système de détection du locuteur, noté $\hat{\alpha}$, comme probabilité *a priori* de présence du locuteur cible, dans le document de test. Ainsi, le score d'un instant t est égal à :

$$\log\left[\frac{s_1(n)}{1-s_1(n)}\right] + \sum_{i=-w}^{w} \log \frac{p(y_{t+i}|X)}{p(y_{t+i}|\Omega)}$$

Une fois ce score instantané calculé, comme pour le premier système de suivi, la décision est prise sur des segments de 5 trames consécutives (50 ms), par comparaison de la moyenne des scores des trames du segment, à un seuil.

Résultats des systèmes de suivi du locuteur et discussion

Pour tous les systèmes de suivi du locuteur, la décision est prise sur le segment, en comparant le score du segment à un seuil. Sur les résultats présentés, le seuil est indépendant du document de test et du locuteur cible.

La figure 4.5 représente les taux d'erreurs de chacun des trois systèmes IRISA, pour différentes valeurs du seuil. Les performances des trois systèmes s'avèrent comparables, mais on peut noter un léger avantage du troisième système (IRISA3).

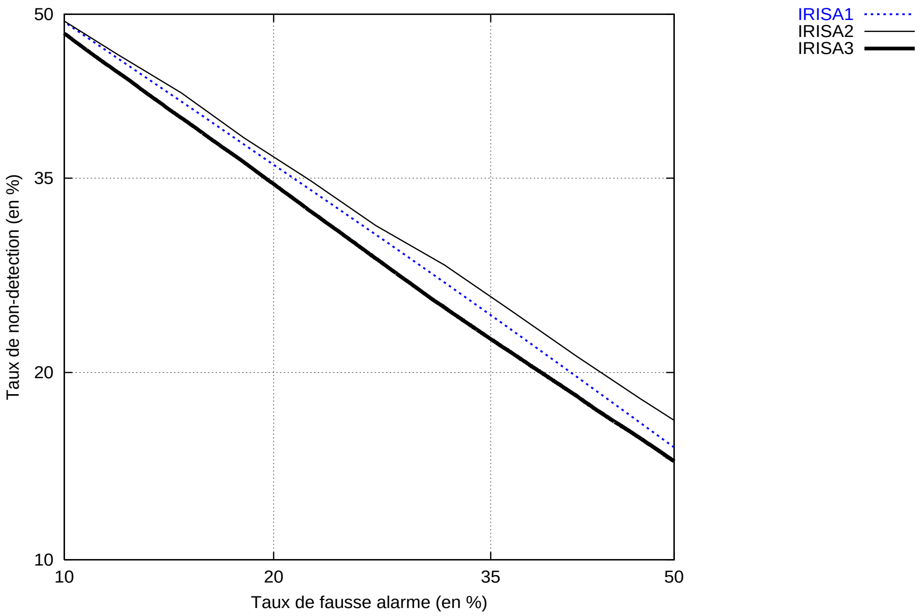

FIG. 4.5 – *Performances des trois systèmes IRISA de suivi du locuteur aux évaluations NIST'99*

Les performances des six systèmes soumis par les laboratoires participant à la tâche de suivi du locuteur (dont le troisième système de l'IRISA), se tiennent sur une fourchette de moins de 5% d'écart.

4.4.5 Conclusion sur les évaluations NIST 1999

Pour cette seconde participation de l'IRISA (au sein du consortium ELISA) aux campagnes d'évaluations NIST en Reconnaissance Automatique du Locuteur, les fonctionnalités de la plaforme ELISA et de nouvelles pistes de recherche ont été explorées.

En ce qui concerne la détection du locuteur, l'utilisation d'une modélisation des vecteurs acoustiques du document de test par un mélange du modèle du locuteur et du modèle du monde, a été introduite. L'estimation du poids du modèle du locuteur cible dans le mélange en question, est proposé comme score du document de test pour la détection du locuteur cible.

4.5 Les systèmes IRISA/ELISA aux évaluations NIST 2000

Pour la campagne d'évaluation NIST 2000, les systèmes soumis par l'IRISA, ont été sensiblement les mêmes que ceux soumis en 1999. Les efforts visant à améliorer l'ordre (le nombre de Gaussiennes) et la structure des covariances des modèles GMM et les essais de nouvelles paramétrisations acoustiques, basées sur les Composantes Discriminantes du Spectre [Magrin-Chagnolleau et al., 2000] n'ont pas conduit à des améliorations des systèmes IRISA de l'année dernière.

La première différence porte sur la paramétrisation acoustique utilisée. Pour l'évaluation 2000, les vecteurs acoustiques sont constitués des 12 coefficients ceptraux, de l'energie (en dB) et des deltas de l'ensemble de ces paramètres. Soit un vecteur de 26 composantes. Les paramètres acoustiques sont calculés avec les mêmes spécifications que celles de l'année 1999, décrits en détail à la section 4.4.1. Ce choix est motivé par le fait que certains des meilleurs systèmes de l'année dernière sont basés sur le même type de coefficients cepstraux et que la nouvelle paramétrisation, comptant moins de coefficients, apporte un gain en temps de calcul pour les expériences de développement.

4.5.1 Détection du locuteur

Deux configurations du second système de détection du locuteur de l'année dernière ont été soumis aux évaluations NIST 1999. La première configuration n'utilise pas de soustraction cepstrale, alors que dans la seconde configuration, une soustraction de la moyenne cepstrale à l'échelle du document est appliquée.

La figure 4.6 représente les courbes DET des deux configurations de système. On constate sur cette figure, que la soustraction cepstrale à l'échelle du document, bien que n'ayant pas de justification théorique, permet d'améliorer les performances du système.

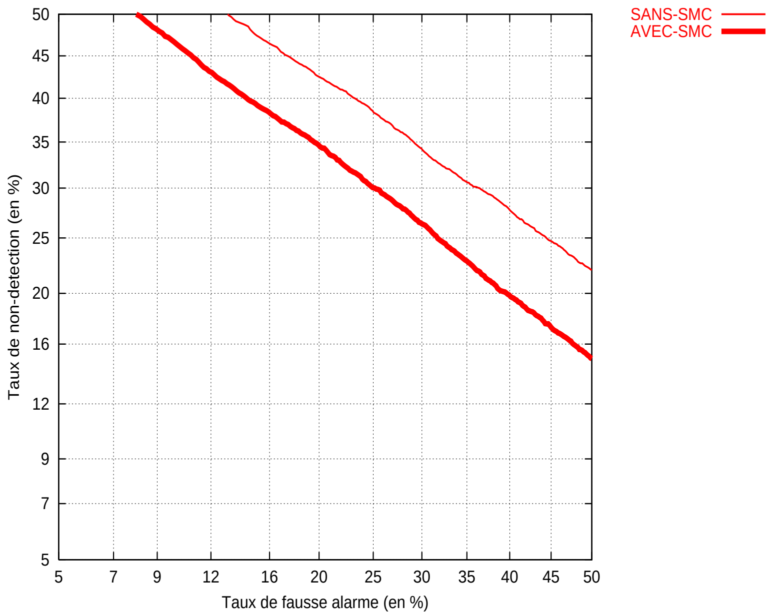

FIG. 4.6 – *Détection du locuteur. Performances du système IRISA, avec et sans Soustraction de la Moyenne Cepstrale.*

Avec environ 30% d'EER, les performances du meilleur système de l'IRISA restent assez éloignées des deux meilleurs systèmes soumis à ces évaluations, qui présentent un EER (Equal Error Rate) entre 15% et 17%. Ce sont par ailleurs, les seuls systèmes dont l'EER est en dessous des 20%.

4.5.2 Suivi du locuteur

Pour le suivi du locuteur, deux systèmes ont été soumis, sous deux configurations de paramétrisations acoustiques différentes, selon qu'une soustraction cepstrale à l'echelle du document, est appliquée ou pas. Les deux systèmes sont basés sur un calcul du score et une décision sur des segments réguliers d'une durée de 0.5 seconde. Le score du segment est calculé sur une fenêtre de lissage d'environ 2 secondes. Les deux systèmes diffèrent seulement par le score utilisé. Le premier système calcule un logarithme du rapport de vraisemblance moyen, utilisant une fenêtre de pondération de Hamming, et le score du second système est le poids du modèle du locuteur cible, dans le modèle de mélange locuteur/monde, estimé sur les vecteurs acoustiques de la fenêtre.

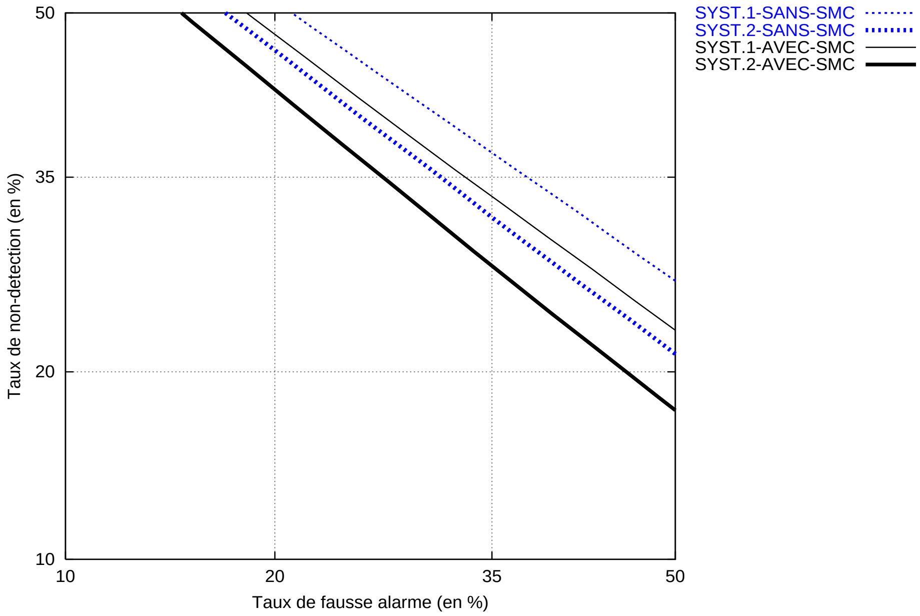

FIG. 4.7 – *Suivi du locuteur. Performances des deux systèmes IRISA, avec et sans soustraction cepstrale.*

La figure 4.7 représente les courbes DET des deux configurations de système. On constate sur cette figure, que la soustraction cepstrale à l'échelle du document, bien que n'ayant pas de justification théorique, permet d'améliorer les performances du système.

Comme pour la tâche de détection, le score basé sur le modèle de mélange donne de meilleures performances, que le log-rapport de vraisemblance moyen. La soustraction cepstrale améliore ces performances dans les deux systèmes.

4.5.3 Conclusions et perspectives

Les résultats des évaluations NIST 2000 ont montré une différence sensible des systèmes IRISA par rapport aux meilleurs systèmes soumis par les autres laboratoires : un taux d'erreur environ 2 fois plus élevé.

D'autres travaux de l'équipe actuellement en cours visent à améliorer la qualité des modèles grâce à des méthodes d'apprentissage basées sur l'approche MAP (Maximum A Posteriori), semblables à celles implémentées par les meilleurs systèmes. Il apparaît également souhaitable de mettre en oeuvre une soustraction cepstrale sur une fenêtre glissante. La prochaine section présente des résultats de travaux complémentaires qui prospecte l'impact de l'utilisation d'une méthode d'apprentissage qui s'apprenente au MAP et d'une soustraction cepstrale glissante.

4.6 Travaux complémentaires

Les principaux aspects sur lesquels les systèmes IRISA/ELISA diffèrent des meilleurs systèmes présentés aux évaluations NIST, sont la modélisation et la paramétrisation acoustique. En s'inspirant des autres systèmes, nous avons conduits quelques expériences complémentaires sur ces deux aspects dans la perspective d'améliorer les performances de nos systèmes.

4.6.1 Travaux complémentaires sur la Modélisation

Sur le plan de la modélisation, les systèmes qui ont obtenu les meilleures performances aux évaluations NIST 2000, utilisent une méthode d'apprentissage du modèle du locuteur, basée sur une adaptation à partir d'un modèle de base appris sur un grand volume de données (qui peut être le modèle du monde). Cette méthode d'apprentissage par adaptation est due à Reynolds, et est présentée par exemple dans [Reynolds et al., 2000]. Elle s'apparente à de l'estimation Bayesienne dans le sens où le modèle de base joue un rôle analogue à celui d'une connaissance *a priori* sur la distribution à estimer. Cependant, elle ne rentre pas exactement dans le formalisme MAP (Maximum A Posteriori) pour l'estimation des MMG. Pour plus de précision sur cette méthode, le lecteur peut consulter les travaux présentés dans [Reynolds et al., 2000], auxquels on se réfère dans cette étude.

Après avoir spécifié la paramétrisation acoustique utilisée, nous présentons des expériences comparant deux types de modèles obtenus par adaptation à partir d'un modèle de base.

Paramétrisation acoustique utilisée

Dans ces expériences, le vecteur de paramètres acoustiques utilisé est constitué de 16 coefficients cepstraux, de leurs deltas et du delta du log-energie : soit un vecteur de 33 composantes. Les coefficients cepstraux sont obtenus de la même manière que ceux utilisés dans les systèmes IRISA/ELISA 2000 (avec une soustraction cepstrale à l'échelle du document[4]).

Lors des précédentes campagnes NIST, des expériences ont montré que l'élimination des plages de silence peut contribuer à l'amélioration des résultats en RAL. Les méthodes de détection automatique de silence les plus utiliséss dans la littérature sont basées sur l'énergie de la trame. Sur le corpus Switchboard, dont sont extraites les données des évaluations NIST, le détecteur de silence utilisé dans [Reynolds et al., 2000] élimine 20 à 25% du signal de parole. La détection de silence n'étant pas la préoccupation majeure de cette étude, nous avons choisi d'éliminer systématiquement 25% des trames de plus faible énergie, pour chaque document (apprentissage ou test).

Résultats de comparaison des différentes modélisations et interprétations

Nous présentons ici des résultats d'expériences dont l'objet est de tester deux modélisations qui procèdent par adaptation à partir d'un modèle de base. Nous rappelons que les modèles du locuteur des systèmes IRISA/ELISA 2000 sont des MMG appris au Maximum de Vraisemblance (MV) par l'algorithme EM. Dans cette partie du travail deux types de modèles adaptés (aussi des MMG à covariances diagonales) sont étudiés : les premiers, notés AD-1024, sont d'ordre 1024 et sont obtenus par adaptation d'un modèle indépendant du genre du locuteur. Les seconds, d'ordre 512, sont obtenus à partir d'un modèle dépendant du genre du

4. Il convient de préciser que cette technique de compensation de canal est bien adaptée à la vérification du locuteur pour lequel seul un canal est présent le document de test. Nous rappelons que cette technique bien que théoriquement inadaptée à la détection du locuteur où le document en question alterne deux canaux, a amélioré les résultats sur les sytèmes IRISA/ELISA 2000

locuteur, et sont alors notés ADG-512.

Nous allons d'abord préciser ci-après les sous-modèles du monde et le modèle du monde utilisés dans ces expériences, puis les modèles de base qui ont servi à construire les modèles adaptés. Par la suite, une première expérience présente les résultats obtenus par les modèles ADG-512 dans une tâche de vérification du locuteur, et la seconde expérience compare les modèles ADG-512 et AD-1024, dans une tâche de détection du locuteur dans une conversation.

Description des différentes modélisations

On rappelle que les quatres sous-modèles du monde dépendants du genre et du type combiné utilisés, sont des MMG d'ordre 256 avec des matrices de covariance diagonales. L'ordre de ces sous-modèles est choisi par rapport aux travaux de [Reynolds et al.]. Pour chaque genre, un sous-modèle du monde d'ordre 512 est constitué en combinant à poids égaux, les deux sous-modèles du monde (d'ordre 256) correspondant aux deux types de combiné. Enfin, un modèle du monde d'ordre 1024 est construit en mélangeant de manière équiprobable les deux sous-modèles du monde dépendants du genre[5]. Les modèles AD-1024 sont adaptés à partir du modèle du monde (d'ordre 1024) et les modèles ADG-512 sont obtenus pour un locuteur donné, par adaptation à partir du sous-modèle du monde d'ordre 512 correspondant au genre du locuteur considéré.

Première expérience

La première expérience a pour objet de situer les performances de l'apprentissage par adaptation à partir d'un modèle de base. La tâche choisie pour cette expérience est la vérification du locuteur. Les performances données correspondent au modèles ADG-512.

5. Ce qui est équivalent à combiner à poids égaux, les 4 sous-modèles du monde (d'ordre 256) dépendants du genre et du type de combiné.

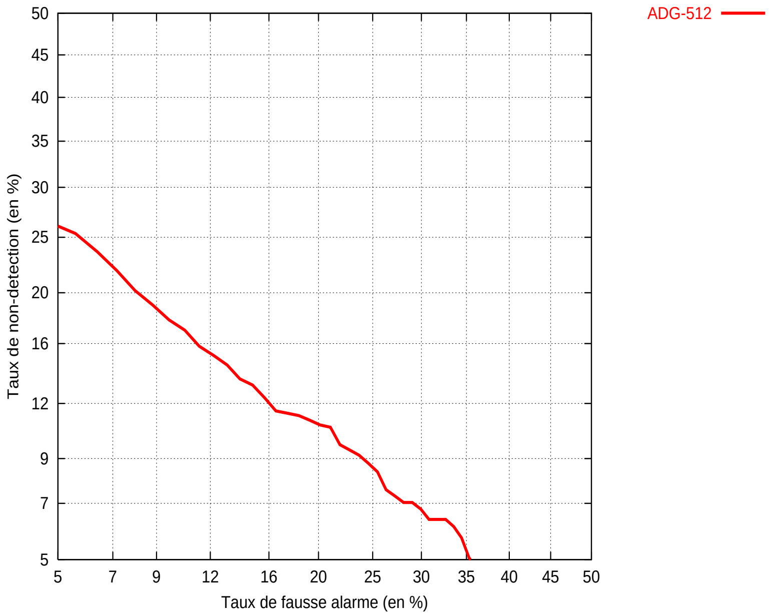

FIG. 4.8 – *Performances du modèle ADG-512 en vérification du locuteur avec utilisation du rapport de vraisemblance. Les tests sont effectués sur un sous-ensemble de* 1061 *documents de tests issus des évaluations NIST* 2000.

Le système de vérification utilisé est un des plus classiques. Il consiste simplement à prendre une décision en comparant à un seuil, le log-rapport de vraisemblance moyen entre les modèles du locuteur et du monde, calculé sur l'ensemble des vecteurs (observations) acoustiques extraits du document de test (sans aucune normalisation). Les tests sont effectués sur un sous-ensemble des données d'évalutations NIST 2000, constitué de 1061 documents de test, dont 505 femmes et 556 hommes. Chaque document est testé contre 11 locuteurs cibles, en utilisant l'ensemble des locuteurs cibles potentiels des évaluations NIST 2000.

La figure 4.8 représente les performances du modèle ADG-512. Ce résultat est comparable à ceux obtenus par les systèmes moyens présentés aux évaluations NIST 2000. Il donne un taux d'EER un peu en dessous du double de celui fournit par les meilleurs systèmes [6]

Seconde expérience

Une seconde expérience est effectuée dans le but de mesurer pour la détection du locuteur dans une conversation, les différences de performances entre les deux types de modèles du locuteur (ADG-512 et AD-1024), appris par adaptation à partir d'un modèle de base.

Pour la tâche de détection, les genres des deux locuteurs du document de test peuvent être différents et sont inconnus du système. Le modèle du monde utilisé est alors le MMG d'ordre 1024 indépendant du genre et du type de combiné. Deux types de modèles de locuteurs sont expérimentés : les modèles AD-1024 motivés par le principe d'adaptation du modèle du locuteur à partir du modèle du monde utilisé par le système, et les modèles ADG-512 qui, adaptés à partir d'un modèle de base dépendant du genre du locuteur, peuvent ainsi accentuer la discrimination par rapport au genre.

Les modèles ADG-512 et AD-1024 sont testés sur un sous-ensemble de 550 documents

6. cette analyse doit cependant être prise avec précaution, car les données de test ne sont pas tout à fait identiques.

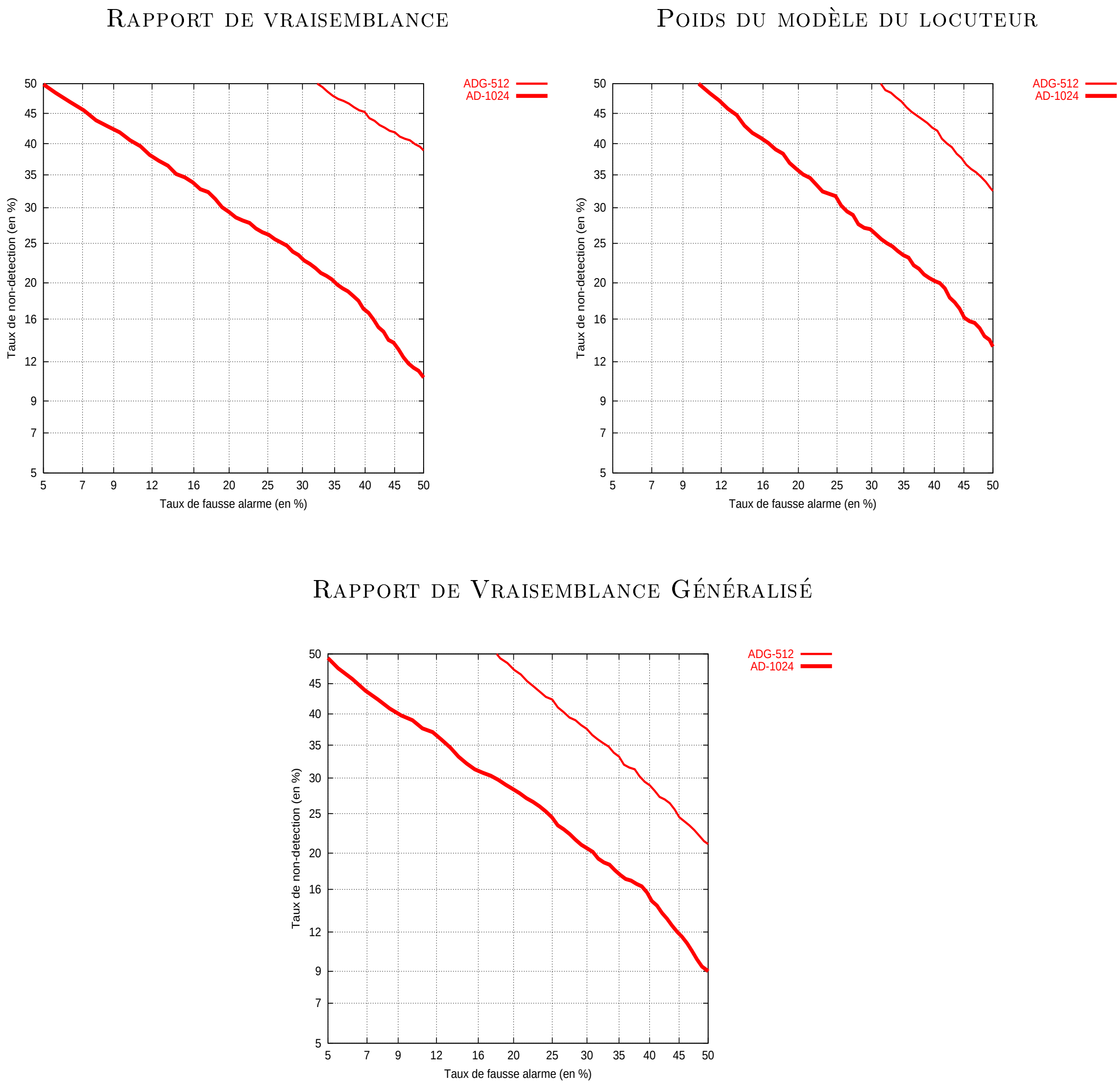

Fig. 4.9 – *Performances des modèles ADG-512 et AD-1024 en vérification du locuteur dans une conversation. Le système de vérification utilisé est basé sur le poids du modèle du locuteur cible dans le mélange locuteur-monde. Les tests sont effectués sur un sous-ensemble de* 1061 *documents des évaluations NIST* 2000.

extraits des données d'évaluations NIST 2000, en sélectionnant un document sur cinq, et en conservant les mêmes locuteurs cibles testés pour chaque document. Les trois scores de détection présentés à la section 2.2.2 sont expérimentés : le rapport de vraisemblance usuel ($s_0(n)$), le poids du modèle du locuteur cible dans le modèle de mélange locuteur-monde ($s_1(n)$) et le rapport de vraisemblance généralisé basé ($s_2(n)$).

La figure 4.9 représente les courbes DET obtenues pour les trois scores avec les deux modèles ADG-512 et AD-1024. On observe nettement, indépendamment du score utilisé, de meilleures performances du modèle AD-1024 par rapport au modèle ADG-512.

Une interprétation possible de ce résultat est la suivante : les modèles ADG-512 étant adaptés à partir d'un modèle de base dépendant du genre du locuteur, ils ont tendance à favoriser les enregistrements sur lesquels les deux locuteurs sont du même genre (où le locuteur cible peut être absent), au détriment de ceux où ils sont de genres différents (où le locuteur peut être présent).

Conclusion sur la modélisation

Nos résultats complémentaires indiquent que, pour la détection du locuteur avec apprentissage par adaptation, le choix du modèle de départ peut avoir une influence déterminante sur la performance du système et qu'il n'est pas forcément judicieux de partir d'un modèle trop spécifique au locuteur.

Par ailleurs, il convient de noter que les performances du système par adaptation restent comparables à celles obtenues aux évaluations NIST, à partir d'un modèle MV à 128 gaussiennes (issu cependant d'une paramétrisation légèrement différente).

4.6.2 Travaux complémentaires sur la paramétrisation

Aux évaluations NIST, les tâches de détection et de suivi du locuteur sont effectuées sur des conversations téléphoniques, sur lesquelles sont présents deux canaux de transmission. On sait que la soustraction de la moyenne cepstrale à l'échelle du document, appliquée jusqu'ici dans nos expériences, est théoriquement inadaptée (comme technique de compensation de canal) à un document où sont présents plusieurs canaux simultanément. Néanmoins, cette technique fait apparaître une amélioration des performances par rapport à une paramétrisation qui n'utilise aucune soustraction cepstrale (voire les résultats des systèmes IRISA/ELISA 2000, section 4.5).

Dans la littérature, une des approches de compensation de canal les plus utilisées pour un document où interviennent plusieurs canaux, est la Soustraction de la Moyenne Cepstrale Glissante (SMCG). L'objet de cette étude est de comparer les performances obtenues par cette méthode à celles d'une Soustraction de la Moyenne Cepstrale classique (SMC).

Dans ces expériences la SMCG est appliquée à titre exploratoire sur une fenêtre d'une durée de 3 s. La figure 4.10 représente les performances des deux types de paramètres acoustiques, en utilisant le proptocole expérimental de détection du locuteur décrit à la section 4.6.1 et les modèles $AD - 1024$. Le système de détection est basé sur le rapport de vraisemblance généralisé.

On observe sur cette figure que la SMCG améliore les performances par rapport à la SMC. Dans nos futurs travaux nos efforts viseront à approfondir ces expériences, notamment sur un réglage des paramètres du filtrage mieux adapaté aux données, et sur la sensibilité des performances à la durèe de la fenêtre d'analyse.

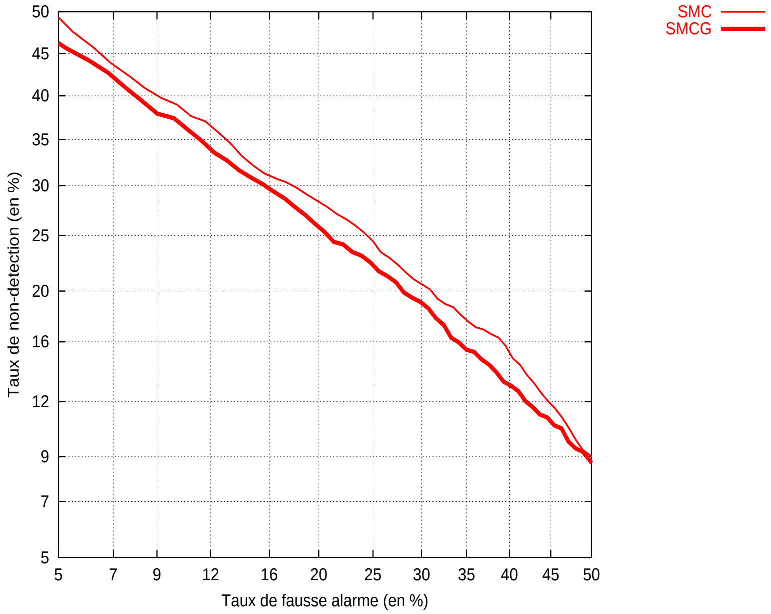

FIG. 4.10 – *Performances en vérification du locuteur de deux types de paramétrisations acoustiques cepstrales : les coefficients obtenus par filtrage RASTA, ceux issus d'une Soustraction de Moyenne Cepstrale à l'échelle du document (SMC) et ceux résultant d'une Soustraction de Moyenne Cepstrale Glissante (SMCG)*

Conclusion générale et perspectives

L'indexation automatique de documents multimédia est un domaine qui suscite actuellement un intérêt considérable et croissant auprès de divers acteurs privés ou publics, scientifiques ou industriels. C'est un domaine relativement récent, où des travaux actuels portent notamment sur la recherche de définitions concises et unifiées des différentes problématiques et schémas d'indexation. Dans ce contexte, cette thèse est une contribution à la définition, l'étude théorique, la mise en œuvre pratique et l'évaluation d'outils automatiques de gestion de bases de documents sonores.

Nous avons défini trois problèmes en vue d'une description simple d'un document sonore : la détection des changements de plage sonore, la détection d'une classe de sons dans un document et le suivi d'une classe de sons dans un document. Ces différents problèmes sont énoncés sous un formalisme très général, de manière à couvrir un champ d'applications relativement vaste.

Dans un souci de généricité, les différents systèmes d'extraction de descripteurs proposés ont une structure modulaire constituée de quatre éléments : la paramétrisation acoustique, la modélisation, le calcul de score et la prise de décision. Les efforts dans cette thèse ont été essentiellement concentrés sur le calcul de score et la prise de décision.

Pour la détection de changements de plage sonore, l'algorithme que nous proposons est basé sur une approche par détection de ruptures statistiques. Alternativement au rapport de vraisemblance généralisé usuel, nous proposons une autre statistique de test d'existence d'une rupture : *le maximum de divergence empirique généralisée.*

Cette dernière statistique est basée sur une estimation de la divergence de Kullback des distributions avant et après rupture. Dans le cas d'une famillle de distributions de type exponentiel, cette statistique est, à un facteur multiplicatif[7] près, égal à la divergence de Kullback entre les distributions estimées avant et après un instant fixé.

L'étude de sa loi asymptotique en absence de rupture a laissé paraître un comportement similaire à celui (connu) du rapport de vraisemblance généralisé. Il convient de continuer dans des travaux ultérieurs à s'intéresser à la comparaison des tests basés sur ces deux statistiques, notamment sur leurs propriétés asymptotiques. De plus, nous faisons appel à une méthode efficace de mise en évidence des extrema dominants d'une statistique de test d'une rupture, calculée à chaque instant.

Les performances fournies par les deux statistiques sur une tâche de détection de changement de plage sonore, correspondant à des ruptures obtenues par concaténation de morceaux de parole et de musique, valident les approches proposées. Cependant, dans un document sonore réel, l'évaluation de cette approche est confrontée à la difficulté à définir un changement de plage sonore de manière concise et objective.

Une des contributions majeures de cette thèse est la proposition d'un modèle de mélange des distributions cible et non-cible, afin de prendre en compte le caractère mixte d'un document où sont simultanément présents des sons cibles et non-cibles. Cette modélisation a donné lieu à deux statistiques pour tester la présence du son cible dans un document : le poids de la distribution cible et le rapport de vraisemblance généralisé.

Sur des expériences en détection du locuteur dans une conversation téléphonique, le poids du modèle cible donne des performances nettement meilleures que celles du rapport de

7. Le nombre d'observations.

vraisemblance usuel. Et l'utilisation du modèle de mélange sur des séquences de vecteurs acoustiques consécutives améliore considérablement les performances. Cette observation suggère d'accentuer l'effort vers une meilleure prise en compte des durées des plages sonores, en utilisant des modèles de durées, comme dans certains travaux sur les Modèles de Markov Cachés.

Les outils de paramétrisation acoustique et de modélisation sont identiques pour tous les systèmes et correspondent à l'état de l'art actuel en Traitement de la Parole et du Son. Certaines de nos expériences ont mis en évidence un plafonnement des performances qui peut être dû à la qualité de l'estimation des modèles de distribution ou à l'inadéquation des paramètres acoustiques utilisés au problème traité. Des efforts doivent être consacrés à ces deux questions spécifiques qui conditionnent les performances globales des systèmes.

De nombreux autres aspects de l'indexation de documents multimedia font l'objet de recherches parallèles et c'est de la réunion de l'ensemble de ces travaux que verront progressivement le jour des systèmes d'indexation de documents multimédia de plus en plus complets à plusieurs égards. Cela devrait également aboutir à une formalisation plus unifiée et plus complète des differents problèmes d'indexation et faciliter l'intégration de toutes les compétences.

Quatrième partie

Annexes

Annexe A

Preuve du théorème 4

Cette annexe donne une preuve du théorème 4, portant sur la loi asymptotique en absence de rupture des statistiques obtenues à partir des divergences de Kullback empiriques généralisées instantanées $\{\Gamma_{2,n}(r),\, 1 \leq r < n\}$, par pondération suivant l'instant r avant de prendre le maximum sur tous les instants.

Nous rappelons l'énoncé de ce théorème et quelques notations et expressions utilisées.

$$\Gamma_{2,n}(r) = n\left[\frac{1}{r}\left(L_{r_-}(\hat{\theta}_{r_-}) - L_{r_-}(\hat{\theta}_{r_+})\right) + \frac{1}{n-r}\left(L_{r_+}(\hat{\theta}_{r_+}) - L_{r_+}(\hat{\theta}_{r_-})\right)\right]$$

où :

$$\hat{\theta}_{r_-} = \arg\max_{\theta\in\Theta}\; p(Y_1,\cdots,Y_r|\theta) \qquad , \qquad \hat{\theta}_{r_+} = \arg\max_{\theta\in\Theta}\; p(Y_{r+1},\cdots,Y_n|\theta)$$

$$L_{r-}(\theta) \;=\; \log p(Y_1,\cdots,Y_r|\theta) \;=\; \sum_{i=1}^{r} \log\left(p(Y_i|\theta)\right)$$

$$L_{r+}(\theta) \;=\; \log p(Y_{r+1},\cdots,Y_n|\theta) \;=\; \sum_{i=r+1}^{n} \log\left(p(Y_i|\theta)\right)$$

La version continue $V_{2,n}$ du processus $\Gamma_{2,n}$ est définie par :

$$V_{2,n}(\tau) = \begin{cases} 0 & \text{si} \quad 0 \leq \tau < \frac{1}{n+1} \\ \Gamma_{2,n+1}([(n+1)\tau]) & \text{si} \quad \frac{1}{n+1} \leq \tau < \frac{n}{n+1} \\ 0 & \text{si} \quad \frac{n}{n+1} \leq \tau \leq 1 \end{cases}$$

Rappel de l'énoncé du théorème 4

On se place sous la suite d'hypothèses d'absence de rupture $H_0(n)$ et les conditions (C1) à (C9) en annexe A.1. Soit q une fonction positive sur]0,1[, croissante au voisinage de 0 et décroissante au voisinage de 1.

Alors :

$$\sup_{0\leq\tau\leq 1} \frac{\tau^2(1-\tau)^2 V_{2,n}(\tau)}{q^2(\tau)} \qquad \textit{converge en loi vers} \qquad \sup_{0\leq\tau\leq 1} \frac{B^{(m)}(\tau)}{q^2(\tau)} \tag{A.1}$$

si et seulement si, il existe une constante $c > 0$ pour laquelle :

$$I_{0,1}(q,c) = \int_0^1 \frac{1}{\tau(1-\tau)} \exp\left(-\frac{cq^2(\tau)}{\tau(1-\tau)}\right) \quad < \quad +\infty \tag{A.2}$$

$B^{(m)}$ est la somme des carrés de m ponts Browniens indépendants réels. Un pont Brownien B sur $[0,1]$ est défini à l'aide d'un mouvement Brownien W par la relation

$$B(\tau) = W(\tau) - \tau W(1) \qquad \forall \ \tau \ \in \ [0,1]$$

La preuve du théorème 4 que nous proposons est inspirée de celle du théorème 3 (section 3.3 de la partie II), présentée dans [Csörgő et Horváth, 1997]. Ces travaux supposent certaines conditions de régularité sur la famille de distributions. La section A.1 reprend les conditions de régularité en question, la section A.2 présente des éléments de preuve du théorème 3 et la section A.3 donne une preuve du théorème 4.

A.1 Conditions de Csörgő et Horváth

Nous reprenons ici les conditions énoncées dans [Csörgő et Horváth, 1997]. Pour cela, on commence par introduire certaines notations relatives au log-vraisemblance et à ses différentielles d'ordres 1, 2 et 3.

$$g(y|\theta) = \log p(y|\theta) \qquad \dot{g}(y|\theta_0) = \frac{\partial g(y|\theta)}{\partial \theta}_{|\theta=\theta_0} \qquad \ddot{g}(y|\theta_0) = \frac{\partial^2 g(y|\theta)}{(\partial\theta)^2}_{|\theta=\theta_0}$$

$$L_{r_-}(\theta) = \sum_{i=1}^{r} g(Y_i|\theta) \qquad \dot{L}_{r_-}(\theta) = \sum_{i=1}^{r} \dot{g}(Y_i|\theta) \qquad \ddot{L}_{r_-}(\theta) = \sum_{i=1}^{r} \ddot{g}(Y_i|\theta)$$

$$L_{r_+}(\theta) = \sum_{i=r+1}^{n} g(Y_i|\theta) \quad \dot{L}_{r_+}(\theta) = \sum_{i=r+1}^{n} \dot{g}(Y_i|\theta) \quad \ddot{L}_{r+}(\theta) = \sum_{i=r+1}^{n} \ddot{g}(Y_i|\theta)$$

$$\dddot{g}(y|\theta_0) = \frac{\partial^3 g(y|\theta)}{(\partial\theta)^3}_{|\theta=\theta_0}$$

Les conditions de régularité de Csörgő et Horváth sont les suivantes :

(C1) Pour tout $i \geq 1$ la distribution de la variable Y_i admet une densité par rapport à une mesure ν σ-finie dans $\mathbb{R}^d$.

(C2) Les densités (distributions) $p(.|\theta_1)$ et $p(.|\theta_2)$ générées par deux paramètres θ_1 et θ_2 différents, sont distinctes pour $\theta_1 \neq \theta_2$.

(C3) Pour tout r, $1 \leq r < n$, on peut trouver d'uniques estimateurs du maximum de vraisemblance, $\hat{\theta}_{r_-}$ et $\hat{\theta}_{r_+}$ vérifiant :

$$\dot{L}_{r_-}(\hat{\theta}_{r_-}) = \sum_{i=1}^{r} \dot{g}(Y_i|\hat{\theta}_{r_-}) = 0 \quad \text{et} \quad \dot{L}_{r_+}(\hat{\theta}_{r_+}) = \sum_{i=r+1}^{n} \dot{g}(Y_i|\hat{\theta}_{r_+}) = 0 \tag{A.3}$$

(C4) Il existe un ouvert $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^m$, contenant $\theta_0^\star$, tel que pour tout $y \in \mathbb{R}^d$ les fonctions :

$$\theta \longmapsto \dot{g}(y|\theta) \quad , \quad \theta \longmapsto \ddot{g}(y|\theta) \quad \text{et} \quad \theta \longmapsto \dddot{g}(y|\theta)$$

existent et sont continus sur Θ_0.

(C5) Il existe des fonctions $M_1(y)$ et $M_2(y)$ telles que : $|\dot{g}(y|\theta)| \leq M_1(y)$, $|\ddot{g}(y|\theta)| \leq M_2(y)$ et $|\dddot{g}(y|\theta)| \leq M_2(y)$, $\forall\ \theta \in \Theta_0$ et $y \in \mathbb{R}^d$.

La notation $|.|$ désigne la norme sup.

M_1 et M_2 satisfont en plus $\int_{\mathbb{R}^d} M_1(y)dy < +\infty$ et $\mathbb{E}_{\theta_0^\star}[M_2(Y)] < +\infty$.

(C6) $\mathbb{E}_\theta \left[\dot{g}(Y|\theta)\right] = 0, \forall\, \theta \in \Theta_0$.

(C7) La matrice d'information de Fisher $I(\theta) = \mathbb{E}_\theta \left[\dot{g}(Y|\theta)\dot{g}^T(Y|\theta)\right]$ et son inverse $I(\theta)^{-1}$ existent et sont continus $\forall\, \theta \in \Theta_0$

(C8) $\mathbb{V}_{\theta_0^\star} \left[\ddot{g}(X|\theta_0^\star)\right] < +\infty$

(C9) Il existe $\mu > 2$ tel que $\mathbb{E}_{\theta_0^\star} \left|\dot{g}(X|\theta_0^\star)\right|^\mu < +\infty$

A.2 Éléments de preuve du théorème 3

Nous commençons par rappeler l'énoncé du théorème 3. Mais avant cela, nous allons d'abord reprendre quelques notations nécessaires à cet énoncé.

$$\begin{aligned}
\Gamma_{1,n}(r) &= \log \frac{\sup\limits_{\theta_- \in \Theta} p(Y_1, \cdots, Y_r|\theta_-) \sup\limits_{\theta_+ \in \Theta} p(Y_{r+1}, \cdots, Y_n|\theta_+)}{\sup\limits_{\theta_0 \in \Theta} p(Y_1, \cdots, Y_n|\theta_0)} \\
&= L_{r_-}(\hat{\theta}_{r_-}) + L_{r_+}(\hat{\theta}_{r_+}) - L_{n_-}(\hat{\theta}_{n_-})
\end{aligned}$$

$$V_{1,n}(\tau) = \begin{cases} 0 & \text{si} \quad 0 \leq \tau < \frac{1}{n+1} \\ 2\Gamma_{1,n+1}([(n+1)\tau]) & \text{si} \quad \frac{1}{n+1} \leq \tau < \frac{n}{n+1} \\ 0 & \text{si} \quad \frac{n}{n+1} \leq \tau \leq 1 \end{cases}$$

Rappel de l'énoncé du théorème 3

On se place sous la suite d'hypothèses $H_0(n)$ et les conditions (C1) à (C9) en annexe A.1. Soit q une fonction positive sur]0,1[, croissante au voisinage de 0 et décroissante au voisinage de 1. Alors :

$$\sup_{0 \leq \tau \leq 1} \frac{\tau(1-\tau)V_{1,n}(\tau)}{q^2(\tau)} \qquad \textit{converge en loi vers} \qquad \sup_{0 \leq \tau \leq 1} \frac{B^{(m)}(\tau)}{q^2(\tau)}$$

si et seulement si, il existe une constante $c > 0$ pour laquelle :

$$I_{0,1}(q,c) = \int_0^1 \frac{1}{\tau(1-\tau)} \exp\left(-\frac{cq^2(\tau)}{\tau(1-\tau)}\right) \quad < \quad +\infty$$

Soient $\mathbf{V}_{r_-}$ et $\mathbf{V}_{r_+}$ les gradients de log-vraisemblances avant et après l'instant r, pour le vrai paramètre de distribution en absence de rupture, $\theta_0^\star$:

$$\mathbf{V}_{r_-} = \dot{L}_{r_-}^T(\theta_0^\star) = \sum_{i=1}^{r} \dot{g}(Y_i|\theta_0^\star)^T \qquad \text{et} \qquad \mathbf{V}_{r_+} = \dot{L}_{r_+}(\theta_0^\star)^T = \sum_{i=r+1}^{n} \dot{g}(Y_i|\theta_0^\star)^T \tag{A.4}$$

On définit la forme quadratique $Q_{1,n}(r)$ en fonction de $\mathbf{V}_{r_-}$, $\mathbf{V}_{r_+}$ et de la matrice d'information de Fisher en $\theta_0^\star$, $I(\theta_0^\star) = I_0^\star$:

$$Q_{1,n}(r) \;=\; \frac{1}{r}\,\mathbf{V}_{r_-}^T\,(I_0^\star)^{-1}\,\mathbf{V}_{r_-} \;+\; \frac{1}{n-r}\,\mathbf{V}_{r_+}^T\,(I_0^\star)^{-1}\,\mathbf{V}_{r_+} \;-\; \frac{1}{n}\,\mathbf{V}_{n_-}^T\,(I_0^\star)^{-1}\,\mathbf{V}_{n_-} \tag{A.5}$$

La preuve du théorème 3 présentée dans [Csörgő et Horváth, 1997] procède en trois étapes :

1. D'abord, par une première phase d'approximation de $2\Gamma_{1,n}(r)$ par la forme quadratique $Q_{1,n}(r)$. (Voire théorème 7 ci-après).
2. Ensuite, dans une deuxième phase, cette approximation est transposée puis affinée sur $V_{1,n}$, la version continue de $2\Gamma_{1,n}$, $V_{1,n}$ définie pour $\tau \in [0,1]$. Plus précisément, pout $\tau \in [0,1]$, $V_{1,n}(\tau)$ est approché par $\frac{B_n^{(m)}}{\tau(1-\tau)}$, avec $B_n^{(m)} = \sum_{i=1}^m B_{n,i}^2$, où les $B_{n,i}$ sont des ponts browniens réels indépendants. (Voire théorème 8 ci-après).
3. Finalement, la troisième phase consiste à déduire le théorème 3 de la seconde phase d'approximation.

Les résultats d'approximations correspondant à la première et à la deuxième phase sont respectivement énoncés par les théorèmes 7 et 8 ci-après.

Théorème 7

Sous la suite d'hypothèses $H_0(n)$ et les conditions $(C1)$ à $(C8)$, pour tout $0 \leq \alpha < \frac{1}{2}$

$$n^\alpha \max_{1\leq r<n} \left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right)^\alpha \mid 2\Gamma_{1,n}(r) \;-\; Q_{1,n}(r)| = \mathcal{O}_p(1) \tag{A.6}$$

et

$$\max_{1\leq r<n} \left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right) \left|2\Gamma_{1,n}(r) \;-\; Q_{1,n}(r)\right| = \mathcal{O}_p(n^{-\frac{1}{2}}(\log\log n)^{\frac{3}{2}}) \tag{A.7}$$

Théorème 8

Sous la suite d'hypothèses $H_0(n)$ et les conditions $(C1)$ à $(C9)$, on peut trouver une suite de processus stochastiques $\{B_n^{(m)}(\tau),\ \tau\ \in\ [0,1]\}$ identiquement distribués selon la loi de $B^{(m)}(\tau) = \sum_{i=1}^m B_i^2(\tau)$, où les B_i sont des ponts browniens réels indépendants, telle que:

pour tout $\lambda > 0$ *et* $0 \leq \alpha < \frac{1}{2} - \frac{1}{\mu}$

$$n^{\alpha} \sup_{\frac{\lambda}{n}\leq\tau\leq 1-\frac{\lambda}{n}} (\tau\,(1-\tau))^{\alpha} \left|V_{1,n}(\tau) \;-\; \frac{B_n^{(m)}(\tau)}{\tau(1-\tau)}\right| = \mathcal{O}_p(1) \tag{A.8}$$

et

$$\sup_{0\leq\tau\leq 1} \left|\tau(1-\tau)V_{1,n}(\tau) \;-\; B_n^{(m)}(\tau)\right| = o_p(n^{(1/\mu)-(1/2)}) \tag{A.9}$$

Compte tenu de A.6, le résultat A.8 est obtenu par approximation de $Q_{1,n}([r\tau])$ par $\frac{B_n^{(m)}(\tau)}{\tau(1-\tau)}$. Plus précisément, on montre que:

$$n^{\alpha} \sup_{\frac{\lambda}{n}\leq\tau\leq 1-\frac{\lambda}{n}} (\tau\,(1-\tau))^{\alpha} \left|Q_{1,n}([r\tau]) \;-\; \frac{B_n^{(m)}(\tau)}{\tau(1-\tau)}\right| = \mathcal{O}_p(1) \tag{A.10}$$

Le résultat A.9 est obtenu de manière similaire en utilisant A.7

A.3 Preuve du théorème 4

Pour une preuve du théorème 4, nous allons montrer qu'asymptotiquement, $\frac{r}{n}(1-\frac{r}{n})\Gamma_{2,n}(r)$ se comporte de la même manière que $2\Gamma_{1,n}(r)$. Cela, en suivant le plan de la preuve du théorème 3, mentionné à la section A.2

Ainsi, comme pour $2\Gamma_{1,n}(r)$, nous effectuons d'abord une approximation de $\frac{r}{n}(1-\frac{r}{n})\Gamma_{2,n}(r)$ par $Q_{1,n}(r)$. Il s'agit des mêmes résultats que ceux énoncés au théorème 7 où on remplace littéralement $2\Gamma_{1,n}$ par $\frac{r}{n}(1-\frac{r}{n})\Gamma_{2,n}$.

Théorème 9

Sous la suite d'hypothèses $H_0(n)$ et les conditions $(C1)$ à $(C8)$, pour tout $0 \leq \alpha < \frac{1}{2}$

$$\max_{1\leq r<n} \left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right)^{\alpha} \left|\left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right) \Gamma_{2,n}(r) \ - \ Q_{1,n}(r)\right| = \mathcal{O}_p(n^{-\alpha}) \tag{A.11}$$

et :

$$\max_{1\leq r<n} \left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right) \left|\left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right) \Gamma_{2,n}(r) \ - \ Q_{1,n}(r)\right| = \mathcal{O}_p(n^{-\frac{1}{2}}(\log\log n)^{\frac{3}{2}}) \tag{A.12}$$

Ensuite, le théorème 4 se déduit du théorème 9 de la même manière que le théorème 3 est obtenu à partir du théorème 7. Pour finir, nous allons maintenant donner une preuve du théorème 9.

Preuve du théorème 9

Le théorème 9 s'appuie sur quelques résultats dus à [Csörgő et Horváth, 1997] que nous reprenons dans les deux lemmes suivants.

Lemme A.3.1

Sous les conditions (C1) à (C7), pour tout $\varepsilon > 0$, on peut trouver $\delta_1 = \delta_1(\varepsilon) > 0$, $T_1 = T_1(\varepsilon)$ et $n_1 = n_1(\varepsilon)$ tels que, pour tous $T \geq T_1$ et $n \geq n_1$:

$$\mathbb{P}\left[\max_{T\leq r\leq n-T} \left(\frac{r}{\log\log r}\right)^{\frac{1}{2}} \left|\hat{\theta}_{r_-} \ - \ \theta_0^\star\right| \ > \ \delta_1\right] \ < \ \varepsilon \tag{A.13}$$

$$\mathbb{P}\left[\max_{T\leq r\leq n-T} \left(\frac{n-r}{\log\log(n-r)}\right)^{\frac{1}{2}} \left|\hat{\theta}_{r_+} \ - \ \theta_0^\star\right| \ > \ \delta_1\right] \ < \ \varepsilon \tag{A.14}$$

Lemme A.3.2

Sous les conditions (C1) à (C7), pour tout $\varepsilon > 0$, on peut trouver $\delta_2 = \delta_2(\varepsilon) > 0$, $T_2 = T_2(\varepsilon)$ et $n_1 = n_1(\varepsilon)$ tels que, pour tous $T \geq T_2$ et $n \geq n_2$:

$$\mathbb{P}\left[\max_{T\leq r\leq n-T}\left(\frac{r}{\log\log r}\right)\left|\hat{\theta}_{r_-} - \theta_0^\star - \mathbf{R}_{r_-}\right| > \delta_2\right] < \varepsilon \tag{A.15}$$

$$\mathbb{P}\left[\max_{T\leq r\leq n-T}\left(\frac{n-r}{\log\log(n-r)}\right)\left|\hat{\theta}_{r_+} - \theta_0^\star - \mathbf{R}_{r_+}\right| > \delta_1\right] < \varepsilon \tag{A.16}$$

où :

$$\mathbf{R}_{r_-} = \frac{1}{r}\,(I_0^\star)^{-1}\,\mathbf{V}_{r_-} \qquad \textit{et} \qquad \mathbf{R}_{r_+} = \frac{1}{n-r}\,(I_0^\star)^{-1}\,\mathbf{V}_{r_+} \tag{A.17}$$

Pour montrer le premier résultat du théorème 9, c'est à dire A.11, on va plus précisément montrer le résultat équivalent :

$$\max_{1\leq r<n}\left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right)^{\alpha+1}\left|\Gamma_{2,n}(r) - Q_{2,n}(r)\right| = \mathcal{O}_p(n^{-\alpha}) \tag{A.18}$$

où :

$$Q_{2,n}(r) = \left(\frac{r}{n}\left(1-\frac{r}{n}\right)\right)^{-1} Q_{1,n}(r) \tag{A.19}$$

$$\begin{aligned}\Gamma_{2,n}(r) - Q_{2,n}(r) \quad &= \quad \frac{n}{r}\left[\left\{L_{r_-}(\hat{\theta}_{r_-}) - L_{r_-}(\hat{\theta}_{r_+})\right\} - \frac{1}{2}\frac{r}{n}Q_{2,n}(r)\right] \\ &\quad + \frac{n}{n-r}\left[\left\{L_{r_+}(\hat{\theta}_{r_+}) - L_{r_+}(\hat{\theta}_{r_-})\right\} - \frac{1}{2}\frac{n-r}{n}Q_{2,n}(r)\right]\end{aligned} \tag{A.20}$$

Remarque : on a $Q_{2,n}(r) = Q_{2,n}(n-r)$

Par analogie entre la première et la deuxième ligne du membre de droite de A.20 ($r \longleftrightarrow n-r$), il suffit de montrer que :

$$\max_{T\leq r<n-T}\left(\frac{r}{n}\right)^{\alpha}\left(1-\frac{r}{n}\right)^{\alpha+1}\left|\left\{L_{r_-}(\hat{\theta}_{r_-}) - L_{r_-}(\hat{\theta}_{r_+})\right\} - \frac{1}{2}\frac{r}{n}Q_{2,n}(r)\right| = \mathcal{O}_p(n^{-\alpha}) \tag{A.21}$$

Un développement de Taylor à l'ordre 2 de $L_{r_-}(\hat{\theta}_{r_+})$ en $\hat{\theta}_{r_-}$, puis de $\ddot{L}_{r_-}(\hat{\theta}_{r_-})$ en θ_0 et la condition (C4) permettent d'écrire :

$$L_{r_-}(\hat{\theta}_{r_-}) - L_{r_-}(\hat{\theta}_{r_+}) = -\frac{1}{2}\,(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T \ddot{L}_{r_-}(\theta_0^\star)(\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) + \sum_{i=1}^{r} U_i \tag{A.22}$$

où :

$$U_i = \mathcal{O}(|\hat{\theta}_{r_-} - \hat{\theta}_{r_+}|^3) + \mathcal{O}(|\hat{\theta}_{r_-} - \hat{\theta}_{r_+}|^2 |\hat{\theta}_{r_-} - \theta_0^\star|)$$

En écrivant $\hat{\theta}_{r_-} - \hat{\theta}_{r_+} = (\hat{\theta}_{r_-} - \theta_0^\star) + (\theta_0^\star - \hat{\theta}_{r_+})$, en développant de manière à faire apparaitre que des puissances de $(\hat{\theta}_{r_-} - \theta_0^\star)$ et de $(\hat{\theta}_{r_+} - \theta_0^\star)$, et en utilisant le lemme A.3.1, on a relativement facilement :

$$\max_{T \le r < n-T} \left(\frac{r}{n}\right)^\alpha \left(1 - \frac{r}{n}\right)^{\alpha+1} \left|\sum_{i=1}^{r} U_i\right| = \mathcal{O}_p(n^{-\alpha}) \tag{A.23}$$

C'est assez fastidieux à effectuer pour tous les termes, nous allons juste le montrer à titre d'exemple pour le terme en $|\hat{\theta}_{r_-} - \theta_0^\star|^3$.

$$\begin{aligned}
\max_{T \le r < n-T} \left(\frac{r}{n}\right)^\alpha \left(1 - \frac{r}{n}\right)^{\alpha+1} r\,\mathcal{O}(|\hat{\theta}_{r_-} - \theta_0|^3) &= \mathcal{O}_p\left(\max_{T \le r < n-T} \left(\frac{r}{n}\right)^\alpha r^{-\frac{1}{2}} (\log\log r)^{-\frac{3}{2}}\right) \\
&= \mathcal{O}_p\left(n^{-\alpha} \max_{T \le r < n-T} r^{\alpha - \frac{1}{2}} (\log\log r)^{-\frac{3}{2}}\right) \\
&= \mathcal{O}_p\left(n^{-\alpha}\right)
\end{aligned}$$

Maintenant, compte tenu de A.22 et A.23, pour obtenir A.21, il suffit de montrer que :

$$\max_{T \le r < n-T} \left(\frac{r}{n}\right)^\alpha \left(1 - \frac{r}{n}\right)^{\alpha+1} \left|(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T \ddot{L}_{r_-}(\theta_0^\star)(\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) + \frac{r}{n} Q_{2,n}(r)\right| = \mathcal{O}_p(n^{-\alpha}) \tag{A.24}$$

Pour cela, on peut d'abord remarquer que :

$$\frac{1}{n}\, Q_{2,n}(r) = (\mathbf{R}_{r_-} - \mathbf{R}_{r_+})^T\, I_0^\star\, (\mathbf{R}_{r_-} - \mathbf{R}_{r_+}) \tag{A.25}$$

Ensuite, on décompose $(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T \ddot{L}_{r_-}(\theta_0^\star)(\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) + \frac{r}{n}\, Q_{2,n}(r)$ en :

$$
\begin{aligned}
&(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T \left(\ddot{L}_{r_-}(\theta_0^\star) + r\, I_0^\star \right) (\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) \\
&- r\, (\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T I_0^\star\, (\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) \;+\; r\, (\mathbf{R}_{r_-} - \mathbf{R}_{r_+})^T\, I_0^\star\, (\mathbf{R}_{r_-} - \mathbf{R}_{r_+})
\end{aligned}
\tag{A.26}
$$

Puis, on s'assure que :

$$
\max_{T \leq r < n-T} \left(\frac{r}{n}\right)^\alpha \left(1 - \frac{r}{n}\right)^{\alpha+1} \left[(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T \left(\ddot{L}_{r_-}(\theta_0) + r I_0^\star \right) (\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) \right] = \mathcal{O}_p\left(n^{-\alpha}\right) \tag{A.27}
$$

et que :

$$
\begin{aligned}
\max_{T \leq r < n-T} \left(\frac{r}{n}\right)^\alpha \left(1 - \frac{r}{n}\right)^{\alpha+1} r\, \left[(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T I_0^\star (\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) - (\mathbf{R}_{r_-} - \mathbf{R}_{r_+})^T I_0^\star (\mathbf{R}_{r_-} - \mathbf{R}_{r_+}) \right] \\
= \mathcal{O}_p\left(n^{-\alpha}\right)
\end{aligned}
\tag{A.28}
$$

Pour montrer A.27, on décompose

$$
(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T \left(\ddot{L}_{r_-}(\theta_0) + r I_0^\star \right) (\hat{\theta}_{r_-} - \hat{\theta}_{r_+})
$$

en :

$$
\begin{aligned}
& (\hat{\theta}_{r_-} - \theta_0^\star)^T \left(\ddot{L}_{r_-}(\theta_0^\star) + r I_0^\star \right) (\hat{\theta}_{r_-} - \theta_0^\star) \\
+2 \quad & (\hat{\theta}_{r_+} - \theta_0^\star)^T \left(\ddot{L}_{r_-}(\theta_0^\star) + r I_0^\star \right) (\hat{\theta}_{r_-} - \theta_0^\star) \\
+ \quad & (\hat{\theta}_{r_+} - \theta_0^\star)^T \left(\ddot{L}_{r_-}(\theta_0^\star) + r I_0^\star \right) (\hat{\theta}_{r_+} - \theta_0^\star)
\end{aligned}
\tag{A.29}
$$

Ensuite, on vérifie que

$$
\max_{T \leq r < n-T} \left(\frac{r}{n}\right)^\alpha \left(1 - \frac{r}{n}\right)^{\alpha+1} [\text{ Chaque ligne de A.29 }] = \mathcal{O}_p\left(n^{-\alpha}\right)
$$

en s'appuyant sur le lemme A.3.1 et le résultat suivant :

$$\max_{1 \leq r < 1} (r \log\log r)^{-\frac{1}{2}} \left| \ddot{L}_{r_-}(\theta_0^\star) + r I_0^\star \right| = \mathcal{O}_p(1)$$

du à la loi du logarithme itéré pour les sommes cumulées.

Le résultat A.28 est obtenu de manière très similaire à A.27. On commence par décomposer

$$(\hat{\theta}_{r_-} - \hat{\theta}_{r_+})^T I_0^\star (\hat{\theta}_{r_-} - \hat{\theta}_{r_+}) - (\mathbf{R}_{r_-} - \mathbf{R}_{r_+})^T I_0^\star (\mathbf{R}_{r_-} - \mathbf{R}_{r_+})$$

en :

$$\begin{aligned} & (\hat{\theta}_{r_-} - \theta_0^\star)^T I_0^\star (\hat{\theta}_{r_-} - \theta_0^\star) - \mathbf{R}_{r_-}^T I_0^\star \mathbf{R}_{r_-} \\ + \quad 2\,[\quad & (\hat{\theta}_{r_-} - \theta_0^\star)^T I_0^\star (\hat{\theta}_{r_+} - \theta_0^\star) - \mathbf{R}_{r_-}^T I_0^\star \mathbf{R}_{r_+} \ ] \\ + \quad & (\hat{\theta}_{r_+} - \theta_0^\star)^T I_0^\star (\hat{\theta}_{r_+} - \theta_0^\star) - \mathbf{R}_{r_+}^T I_0^\star \mathbf{R}_{r_+} \end{aligned} \tag{A.30}$$

Ensuite, on vérifie à l'aide du lemme A.3.2 que

$$\max_{T \leq r < n-T} \left(\frac{r}{n}\right)^\alpha \left(1 - \frac{r}{n}\right)^{\alpha+1} r \ [\text{ Chaque ligne de A.30 }] = \mathcal{O}_p\left(n^{-\alpha}\right)$$

En récapitulant, les résultats A.26, A.27 et A.28 montrent A.24 qui, combiné avec A.22 et A.23, induisent A.21. Enfin, les résultats A.21 et A.20 permettent d'obtenir A.18 qui est équivalent à A.11 du théorème 9.

Le résultat A.12 est obtenu avec des arguments similaires à ceux utilisés pour montrer A.11. À ce titre, sa preuve n'est pas présentée.

Annexe B

Lois asymptotiques de $S_1(n)$ et $S_2(n)$ sous H_0

B.1 Preuve du théorème 5

Le résultat 2.18 (cas $\theta = 1$) se déduisant par analogie de 2.17 (cas $\theta = 0$), en considérant $\theta' = 1 - \theta$, le cas $\theta = 1$ ne sera pas explicitement étudié.

On définit la fonction de vraisemblance L_n par :

$$\theta \longmapsto L_n(\theta) = \sum_{i=1}^{n} \log p_\theta(Y_i) \tag{B.1}$$

Nous rappelons que la statistique $S_1(n)$ est l'estimateur du maximum de vraisemblance noté $\hat{\theta}_n$:

$$S_1(n) = \hat{\theta}_n = \underset{\theta \in [0,1]}{\arg\max}\ L_n(\theta) \tag{B.2}$$

La preuve s'appuie essentiellement sur une méthode classique de recherche de la loi asymptotique d'un estimateur du maximum de vraisemblance.
Cette méthode (voire par exemple la section 2.6.3 de [Genot-Catalot et Picard, 1993]) suppose que la fonction $\theta \longmapsto L_n(\theta)$ est de classe C^2, et consiste dans un premier temps à montrer que $\hat{\theta}_n$ est consistant et vérifie $L'_n(\hat{\theta}_n) = 0$. Ensuite, elle utilise un développement de Taylor à

l'ordre 1 de L'_n autour de la vraie valeur θ_0, calculé en $\hat{\theta}_n$. Soit :

$$0 = L'_n(\hat{\theta}_n) = L'_n(\theta_0) + (\hat{\theta}_n - \theta_0)L''_n(\theta_0) + r_n(\hat{\theta}_n - \theta_0) \tag{B.3}$$

avec :

$$r_n(\hat{\theta}_n - \theta_0) = o(|\hat{\theta}_n - \theta_0|)$$

Ce qui peut encore s'exprimer :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\frac{1}{\sqrt{n}}L'_n(\theta_0)}{\frac{1}{n}[L''_n(\theta_0) + \frac{r_n(\hat{\theta}_n - \theta_0)}{\hat{\theta}_n - \theta_0}]} \tag{B.4}$$

On montre ensuite (sous certaines conditions) que :

$$\frac{1}{\sqrt{n}}L'_n(\theta_0) \qquad \text{converge en loi vers} \qquad \mathcal{N}(0,I(\theta_0)) \tag{B.5}$$

$$\frac{1}{n}L''_n(\theta_0) \qquad \text{converge presque sûrement vers} \qquad I(\theta_0) \tag{B.6}$$

$$\frac{1}{n}\frac{r_n(\hat{\theta}_n - \theta_0)}{\hat{\theta}_n - \theta_0} \qquad \text{converge en probabilité vers} \qquad 0 \tag{B.7}$$

Les formulent B.4, B.5, B.6 et B.7 impliquent l'efficacité de $\hat{\theta}_n$, c'est à dire que :

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0,I(\theta_0)^{-1}) \tag{B.8}$$

Maintenant, dans notre cas, la fonction de vraisemblance L_n est de classe C^2. Sa dérivée seconde est donnée par :

$$L''_n(\theta) = -\sum_{i=1}^{n}\left[\frac{p_1(Y_i) - p_0(Y_i)}{p_\theta(Y_i)}\right]^2 \tag{B.9}$$

Examinons la consistance de $\hat{\theta}_n$. Pour cela, on commence par remarquer que la fonction de vraisemblance L_n est continue sur le compact [0,1], donc $\hat{\theta}_n$ existe. Ensuite, comme la famille de distributions $\{P_\theta;\ 0 \leq \theta \leq 1\}$ est identifiable, pour prouver la consistance de $\hat{\theta}_n$, il suffit de montrer que pour tout θ et θ' dans [0,1], $\mathbb{E}_{P_\theta}\left[\left|\log\frac{p(Y_1|\theta)}{p(Y_1|\theta')}\right|\right] < +\infty$. Cela résulte des hypothèses:

$$\mathbb{E}_{P_\theta}\left[\left|\log\frac{p_\theta(Y_1)}{p_{\theta'}(Y_1)}\right|\right] < \max\left\{\mathbb{E}_{P_0}\left[\left|\log\frac{p_1(Y_1)}{p_0(Y_1)}\right|\right], \mathbb{E}_{P_1}\left[\left|\log\frac{p_1(Y_1)}{p_0(Y_1)}\right|\right]\right\}$$

La consistance de $\hat{\theta}_n$ est alors assurée.

Dans un premier temps, on suppose que $\hat{\theta}_n$ satisfait $L'_n(\hat{\theta}_n) = 0$. Sous cette condition, on montre assez facilement que les résultats B.5, B.6 et B.7 sont vérifiés.
Le résultat B.5 est dû au Théorème de la Limite Centrale, car la variable $\frac{\partial}{\partial\theta} \log p_\theta(Y_1)\big|_{\theta=\theta_0}$ est centrée sous P_{θ_0}, et $I(\theta_0) \in]0, +\infty[$. Le résultat B.6 est obtenu grâce à la Loi Forte des Grands Nombres (car $I(\theta_0) < +\infty$). Quant à B.7, comme la fonction $\theta \longmapsto L_n(\theta)$ a une dérivée troisième continue, donc bornée sur le compact [0,1], le reste de Taylor $r_n(\hat{\theta}_n - \theta_0) = o((\hat{\theta}_n - \theta_0)^2)$. La consistance de $\hat{\theta}_n$ permet alors d'avoir B.7. D'où le résultat B.8.

Dans un second temps, on examine les conditions dans lesquelles on a : $L'_n(\hat{\theta}_n) = 0$? De manière générale, voire par exemple [Titterington et al., 1985] section 4.3 :

$$\begin{array}{lll} \text{si} & L'_n(0) < 0 & , \quad \hat{\theta}_n = 0 \\ \text{si} & L'_n(1) > 0 & , \quad \hat{\theta}_n = 1 \\ \text{sinon} & L'_n(\hat{\theta}_n) = 0 & , \quad 0 < \hat{\theta}_n < 1 \end{array}$$

Si $\theta_0 \in]0,1[$

$$\left\{ \begin{array}{lll} \lim_{n\to+\infty} \mathbb{P}_{\theta_0}\left[L'_n(0) < 0\right] & = & 0 \\ \lim_{n\to+\infty} \mathbb{P}_{\theta_0}\left[L'_n(1) > 0\right] & = & 0 \end{array} \right.$$

Donc $\mathbb{P}_{\theta_0}\left[\hat{\theta}_n \in]0,1[\text{ et } L'_n(\hat{\theta}_n) = 0\right] \longrightarrow 1$. Cela permet de conclure B.8 ou 2.16.
Sous la loi P_0 $(\theta_0 = 0)$:

$$\left\{ \begin{array}{lll} \lim_{n\to+\infty} \mathbb{P}_0\left[L'_n(0) < 0\right] & = & \frac{1}{2} \\ \lim_{n\to+\infty} \mathbb{P}_0\left[L'_n(1) > 0\right] & = & 0 \end{array} \right.$$

On en déduit qu'asymptotiquement, $\mathbb{P}_0\left[L'_n(0) = 0\right] \longrightarrow \frac{1}{2}$. D'ou le résultat 2.17.

B.2 Conditions de régularités d'Ibragimov et Has'minskii

Cette section présente les conditions que doit satisfaire un modèle de densités dit *régulier*, énoncées dans [Genot-Catalot et Picard, 1993] et qui sont reprises de

[Hibragimov et Has'minskii, 1981].

1. Le modèle de probabilités est dominé par une mesure μ et les densités $p(y|\theta)$ sont mesurable par rapport à μ.
2. La fonction $\theta \longmapsto \frac{\partial}{\partial\theta} p^{\frac{1}{2}}(y,\theta)$ est absolument continue pour μ-presque tout y;
3. La matrice d'information de Fisher

$$I(\theta) = 4\mathbb{E}_\theta \left[\frac{1}{p(Y_1|\theta)} \left(\frac{\partial}{\theta} p^{\frac{1}{2}}(Y_1|\theta) \right)^2 \right]$$

   est finie et définie positive pour tout θ in Θ. La fonction $\theta \longmapsto I(\theta)$ est continue.
4. Si

$$B_{\varepsilon\delta} = \left\{ y, \left| 1 - \sqrt{\frac{p(y|\theta+\varepsilon)}{p(y|\theta)}} \right| > \delta \right\}$$

   alors

$$\sup_{\theta - \theta_0| < \varepsilon} \mathbb{P}_\theta[B_{\varepsilon\delta}] \; = \; o(\varepsilon^2)$$

   pour tout $\theta_0 \in \theta$.
5. Il existe $q > 0$ tel que

$$\sup_{\theta\in\Theta} \frac{I(\theta)}{1 + |\theta|^q} < +\infty$$

6. Il existe $r > 0$ tel que

$$\sup_{\theta\in\Theta} |\theta - \theta_0|^r \mathbb{E}_{\theta_0} \left[\left(\frac{p(Y_1|\theta)}{p(Y_1|\theta_0)} \right)^2 \right]$$

   pour tout $\theta_0 \in \Theta$.

Bibliographie

[Aigrin, 1999] Aigrin, P. (1999). Multimedia content representation : are we tackling the actual problems ? *Proceedings of the first European workshop on Content-Based Multimedia Indexing CBMI'99*, pages 3–5.

[Akaïke, 1974] Akaïke, H. (1974). A new look at statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723.

[Alder, 1997] Alder, M. D. (1997). *An introduction to pattern recognition : statistical, neural net and syntactic methods of getting robots to see and hear.* http://ciips.ee.uwa.edu.au/mike/PatRec/.

[André-Obrecht, 1993] André-Obrecht, R. (1993). Segmentation et parole ? Document d'habilitation.

[Anily et Federgruen, 1968] Anily, S. et Federgruen, A. (1968). Simulated annealing methods with general acceptance probability. *Journal of Applied Probabilities*, 24:657–667.

[Azencott, 1988] Azencott, R. (1988). Simulated annealing. *Séminaire Bourbaki*, 697:1–15.

[Baker, 1975] Baker, J. (1975). The dragon system: an overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(1):29–29.

[Basseville et Benveniste, 1983a] Basseville, M. et Benveniste, A. (1983a). Design and comparative study of some sequential jump detection algorithms for digital signals. *IEEE Transactions on Acoustics, Speech, Signal Processing*, IASSP-31(3):521–535.

[Basseville et Benveniste, 1983b] Basseville, M. et Benveniste, A. (1983b). Sequential detection of abrupt changes in spectral charactaristics of digital signals. *IEEE Transactions on Information Theory*, IT-29(5):709–724.

[Basseville et Nikiforov, 1993] Basseville, M. et Nikiforov, I. V. (1993). *Detection of abrupt changes : theory and application.* PTR Prentice-Hall.

[Beigi et Maes, 1998] Beigi, H. et Maes, S. (1998). Speaker, channel and environment change detection. *World congress of automation.*

[Berrut, 1997] Berrut, C. (1997). Indexation de données multimédia. Document d'habilitation.

[Bertrand, 2000] Bertrand, P. (2000). A local method for estimating change points : the "hat-function". *Statistics*, 34:215–235.

[Bhattacharya et Brockwell, 1976] Bhattacharya, P. K. et Brockwell, P. J. (1976). The minimum of an additive process with applications to signal estimation and storage theory. *Zeit. Wahr. und veriv. Gebiete*, 37:51–75.

[Boëffard-Dosierre, 1993] Boëffard-Dosierre, O. (1993). *Segmentation automatique d'unités acoustiques pour la synthèse de la parole.* Doctorat de traitement du signal, Université de Rennes 1.

[Borko et Bernier, 1978] Borko, H. et Bernier, C. L. (1978). *Indexing concepts and methods.* New York, Academic Press.

[Boves, 1998] Boves, L. (1998). Commercial applications of speaker verification : overview and critical success factors. *Conférence sur la Reconnaissance Automatique du Locuteur et ses Applications Commerciales et Criminalistiques (RLA2C)*, pages 150–159.

[Brandt, 1983] Brandt, A. V. (1983). Detecting and estimating parameter jumps using ladder algorithms and likelihood ratio tests. in *ICASSP*, pages 1017–1020.

[Breiman, 1992] Breiman, L. (1992). *Probability.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics - Philadelphia.

[Brodsky et Darkhovsky, 1993] Brodsky, B. E. et Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems.* Mathematics and its Applications. Kluwer Academic Publishers.

[Calliope, 1989] Calliope (1989). *La parole et son traitement automatique.* Masson, Paris.

[Carlin et al., 1992] Carlin, B. P., Gelfand, A. E., et Smith, A. F. M. (1992). Hierarchical bayesian analysis of change-point problems. *Applied Statistics*, 41(2):389–405.

[Celeux, 1987] Celeux, G. (1987). Reconnaissances de mélanges de densités de probabilité et application à la validation des résultats en classification. Thèse de Doctorat es Sciences Mathématiques.

[Celleux et Diebolt, 1992] Celleux, G. et Diebolt, J. (1992). A stochastic approximation type em algorithm for the mixture problem. *Stochastics and Stochastics Reports*, 41:119–134.

[Chen et Gopalakrishnan, 1998] Chen, S. et Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. *DARPA speech workshop*.

[Childers et al., 1969] Childers, D. G., Skinner, D. P., et C, K. R. (1969). The cepstrum : a guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443.

[Csörgő et Horváth, 1997] Csörgő, M. et Horváth, L. (1997). *Limit theorems in change-point analysis*. Chichester, wiley édition.

[D'Alessandro et Demars, 1991] D'Alessandro, C. et Demars, C. (1991). Représentations temps-fréquence du signal de parole. *Traitement du Signal*, pages 153–173.

[Davis et Mermelstein, 1980] Davis, S. B. et Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing*.

[Delacourt, 2000] Delacourt, P. (2000). *La segmentation et le regroupement par locuteurs pour l'indexation de documents audio*. Thèse de doctorat, Institut Eurecom.

[Delacourt et al., 1999] Delacourt, P., Krize, D., et Wellekens, C. J. (1999). Speaker based segmentation for audio data indexing. *ESCA workshop : accessing information in audio data*.

[Delacourt et Wellekens, 1999] Delacourt, P. et Wellekens, C. J. (1999). A first step into speaker-based indexing. *Proceedings of the first European workshop on Content-Based Multimedia Indexing CBMI'99*, pages 89–95.

[Delyon et al., 1988] Delyon, B., Andre-Obrecht, R., et Su, H. Y. (1988). Expériences en vue du décodage acoustico phonétique à partir d'une recherche d'événements acoustiques et d'un codage vectoriel. *Journal of Acoustique*, 1:229–201.

[Dembo et Zeitouni, 1993] Dembo, A. et Zeitouni, O. (1993). *Large deviations techniques an applications*. Jones and Barlett Books in Mathematics. Jones and Barlett Publishers.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Ser. B*, 39:1–38.

[Deshayes et Picard, 1983] Deshayes, J. et Picard, D. (1983). *Ruptures de modèles en statistique*. Thèse d'état, Université Paris-Sud.

[Deshayes et Picard, 1984a] Deshayes, J. et Picard, D. (1984a). Lois asymptotiques des tests et estimateurs de rupture dans un modèle statistique classique. Technical Report 84C001, E.N.S.T.

[Deshayes et Picard, 1984b] Deshayes, J. et Picard, D. (1984b). Principe d'invariance sur le processus de vraisemblance. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 20(1).

[Deshayes et Picard, 1986] Deshayes, J. et Picard, D. (1986). Off-line statistical analysis of change-point models using nonparametric and likelihood methods. in *Detection of abrupt ahanges in signals and dynamical systems*, Lecture Notes in Control and Information Sciences, pages 103–168. Springer-Verlag.

[Diebolt et Robert, 1991] Diebolt, J. et Robert, C. P. (1991). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society B*, 56(2):363–375.

[DSPJ, 2000] DSPJ, éd. (2000). *Digital Signal Processing*, volume 10. Academic Press. Special Issue: NIST 1999 Speaker Recognition Workshop.

[Dunn et al., 2000] Dunn, R. B., Reynolds, D. A., et Quatieri, T. F. (2000). Approaches to speaker detection and tracking in conversational speech. *Digital Signal Processing*, 10(1-3):93–112. Special Issue: NIST 1999 Speaker Recognition Workshop.

[ELISA, 2000] ELISA (2000). The ELISA systems for the NIST'99 evaluation in speaker detection and tracking. *Digital Signal Processing*, 10(1-3):143–153. Special Issue: NIST 1999 Speaker Recognition Workshop.

[ELISA, site] ELISA (site). *http://www.lia.univ-avignon.fr/equipes/RAL/elisa/index.html.* ELISA Consortium.

[Everitt et Hand, 1981] Everitt, B. S. et Hand, D. J. (1981). *Finite Mixture Distributions.* Monographs on applied probability and statistics. Chapman and Hall.

[Feder, 1975] Feder, P. I. (1975). The log-likelihood ratio in segmented regression. *Annals of Statistics*, 3:84–97.

[Flanagan, 1972] Flanagan, J. L. (1972). *Speech analysis, synthesis and perception.* Springer-Verlag, Berlin.

[Fredouille, 2000] Fredouille, C. (2000). *Approche statistique pour la reconnaissance automatique du locuteur : informations dynamiques et normalisation bayesienne des vraisemblances.* Thèse de doctorat en informatique, Université d'Avignon et des Pays de Vaucluse.

[Gauvain et al., 1999] Gauvain, J.-L., Lamel, L., et Adda, G. (1999). Audio partitionning and transcription for broadcast data indexation. *Proceedings of the first european workshop on Content-Based Multimedia Indexing CBMI'99*, pages 67–73.

[Gauvain et Lee, 1994] Gauvain, J.-L. et Lee, C.-H. (1994). Maximum *a posteriori* estimation for eultivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

[Genot-Catalot et Picard, 1993] Genot-Catalot, V. et Picard, D. (1993). *Éléments de statistique asymptotique.* Springer-Verlag.

[Girshick et Rubin, 1952] Girshick, M. A. et Rubin, H. (1952). A bayes approach to a quality control model. *Annals of Mathematical Statistics*, 23:114–125.

[Gombay et Horváth, 1990] Gombay, E. et Horváth, L. (1990). Asymptotic distributions of the maximum likelihood tests for change in the mean. *Biometrika*, 77(2):411–414.

[Hathaway, 1991] Hathaway, R. J. (1991). A constrained EM algorithm for univariate normal mixtures. *IEEE Transactions on Information Theory*, 37(6):1683–1690.

[Hermansky et Morgan, 1994] Hermansky, H. et Morgan, N. (1994). Rasta-plp speech analysis technique. *International Conference on Acoustics, Speech, and Signal processing*, 2(4):578–589.

[Hermansky et al., 1992] Hermansky, H., Morgan, N., Bayya, N., et Kohn, P. (1992). Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, pages 121–124.

[Herrera et Serra, 1999] Herrera, P. et Serra, X. (1999). A proposal for the description of audio in the context of mpeg-7. *Proceedings of the first European workshop on Content-Based Multimedia Indexing CBMI'99*, pages 81–88.

[Hibragimov et Has'minskii, 1981] Hibragimov, I. A. et Has'minskii, R. Z. (1981). *Statistical estimation, asymtotic theory.* Springer-Verlag.

[Hinkley, 1970] Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(7):1–17.

[Hinkley, 1971] Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523.

[Holst et Lindgren, 1991] Holst, U. et Lindgren, G. (1991). Recursive estimation in mixture models with markov regime. *IEEE Transactions on Information Theory*, 37(6):1683–1690.

[Jelinek, 1976] Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

[Kedem, 1986] Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493.

[Khalil et Duchêne, 1999] Khalil, M. et Duchêne, J. (1999). Une approche pour la détection fondée sur une somme cumulée dynamique associée à une décomposition multiéchelle. application à l'emg utérin. *Grestsi'99*, 2:1049–1052.

[Kiefer, 1959a] Kiefer, J. (1959a). A functionnal equation technique for obtaining Wiener process probabilities associated with theorems of Kolmogorov-Smirnov type. *Math. Proc. Cambridge Phil. Soc.*, 55:328–332.

[Kiefer, 1959b] Kiefer, J. (1959b). K-sample analogues of the Kolmogorov-Smirnov and Cramèr-von Misses tests. *Ann. Math. Stat.*, 30:420–447.

[Kiefer et Wolfowitz, 1956] Kiefer, J. et Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27:887–906.

[LDC, site] LDC (site). *http://morph.ldc.upenn.edu.* Linguistic Data Consortium.

[Lehmann, 1959] Lehmann, E. L. (1959). *Testing statistical hypothesis.* Wiley Publications in Statistics.

[Lehmann, 1961] Lehmann, E. L. (1961). *Testing statistical hypothesis.* Wiley Publications in Statistics, Second edition.

[Lehmann, 1991] Lehmann, E. L. (1991). *Theory of point estimation.* Wadsworth & Brooks, Pacific Grove, California.

[Lienard, 1977] Lienard, J. S. (1977). *Les processus de la communication parlée.* Masson, Paris.

[Magrin-Chagnolleau et al., 2000] Magrin-Chagnolleau, I., Gravier, G., Seck, M., Boëffard-Dosierre, O., Blouet, R., et Bimbot, F. (2000). A further investigation in speech features for sopeaker characterization. *International Conference of Spochen Language Processing (ICSLP).*

[Magrin-Chagnolleau, 1997] Magrin-Chagnolleau, I. (1997). *Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte.* PhD thesis, École Nationale Supérieure des Télécommunications. ENST 97 E 010.

[Mammone et al., 1996] Mammone, R. J., Ramachandra, R. P., et Zhang, X. (1996). Robust speaker recognition : a feature-based approach. *IEEE Signal Processing Magazine*, pages 58–71.

[Martin et al., 1997] Martin, A., Doddington, G., Ordowski, M., et Przybocki, M. (1997). The DET curve in assessment of detection task performance. *Eurospeech'97*, 4:1895–1898.

[Martin et Przybocki, 2000] Martin, A. et Przybocki, M. (2000). The NIST 1999 speaker recognition evaluation : an overview. *Digital Signal Processing*, 10(1-3):1–18. Special Issue: NIST 1999 Speaker Recognition Workshop.

[Matei et al., 1999] Matei, M., Delfosse, N., Staroswiecki, M., et Darkhovski, M. (1999). Une nouvelle approche pour la détection en présence de paramètres inconnus. *Grestsi'99*, 2:981–984.

[McLachlan et Basford, 1988] McLachlan, G. J. et Basford, K. E. (1988). *Mixture models: inference and applications to clustering*, volume 84 of *Statistics: Textbooks and Monographs.* Marcel Dekker Inc.

[Meignier et al., 2000] Meignier, S., Bonastre, J.-F., Fredouille, C., et Merlin, T. (2000). Evolutive HMM for multi-speaker tracking system. *International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[Montacié et Caraty, 1997] Montacié, C. et Caraty, M.-J. (1997). Sound channel video indexing. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 2359–2362.

[Moon, 1996] Moon, T. K. (1996). The Expectation-Maximisation algorithm. *IEEE Signal Processing Magazine*, pages 47–60.

[Moustakides, 1986] Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distribution. *The Annals of Statistics*, 14(4):1479–1387.

[Nikiforov, 1995] Nikiforov, I. V. (1995). A generalized change detection problem. *IEEE Transactions on Information Theory*, 41(1):171–187.

[Nikiforov, 1997] Nikiforov, I. V. (1997). Two strategies in the problem of change detection and isolation. *IEEE Transactions on Information Theory*, 43(2):770–776.

[NIST, site] NIST (site). *http://www.nist.gov/speech.* National Institute of Standards and Technologies.

[Ò Ruanaidh et Fitzgerald, 1996] Ò Ruanaidh, J. J. K. et Fitzgerald, W. J. (1996). *Numerical bayesian methods applied to signal processing*, chapter 5, pages 96–119. Statistics and Computing. Springer-Verlag.

[Oppenheim, 1969] Oppenheim, A. V. (1969). Speech analysis-synthesis based on homomorphique filtering. *J.A.S.A.*, 45(2):458–465.

[O'Shaughnessy, 1987] O'Shaughnessy, D. (1987). *Speech communication : human and machine.* Ma: Addison-Wesley Publishing Company, 1 ed. Reading.

[Page, 1954] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 4:100–115.

[Page, 54b] Page, E. S. (54b). Control charts for the mean of a normal population. *Journal of the Royal Statistical Society*, 16(1):136–139.

[Pollak, 1985] Pollak, M. (1985). Optimal detection of a change in distribution. *The Annals of Statistics*, 13(1):206–227.

[Quandt, 1960] Quandt, R. E. (1960). Tests of the hypothesis that a linear system obeys two separate regimes. *J.A.S.A.*, 55:324–330.

[Rabiner et Juang, 1993] Rabiner, L. R. et Juang, B.-H. (1993). *Fundamentals of speech recognition.* Prentice-Hall.

[Redner et Walker, 1984] Redner, R. A. et Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239.

[Reynolds, 1992] Reynolds, D. A. (1992). *A gaussian mixture modeling approach to text-independent speaker identification.* Phylosophy doctor thesis in electrical engineering, Georgia Institute of Technology.

[Reynolds et al., 2000] Reynolds, D. A., Dunn, R. B., et Quatieri, T. F. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41. Special Issue: NIST 1999 Speaker Recognition Workshop.

[Ridolfi et Idier, 1999] Ridolfi, A. et Idier, J. (1999). penalized maximum likelihood estimation for univariate normal mixture distributions. *Grestsi'99*, 2:259–262.

[Rissanen, 1978] Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

[Roberts et al., 1998] Roberts, S. J., Husmeier, D., Rezek, I., et Penny, W. (1998). Bayesian approches to gaussian mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11).

[Rowley, 1988] Rowley, J. E. (1988). *Abstracting and indexing (2nde edition).* Londres, Clive Bingle.

[Salton et G, 1983] Salton, G. et G, M. J. (1983). *Introduction to modern information retrieval.* McGraw Hill Book Company, New York.

[Saporta, 1990] Saporta, G. (1990). *Probabilités, analyse de données et statistique.* Technip.

[Saunders, 1996] Saunders, J. (1996). Real-time discrimination of broadcast speech/music. *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2:993–996.

[Scheirer et Slanley, 1997] Scheirer, E. et Slanley, M. (1997). Construction and evaluation of a multifeature speech/music discriminator. *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2:1331–1334.

[Seck et al., 1998] Seck, M., Bimbot, F., et Delyon, B. (1998). Classification parole/musique de signaux sonores paramétrisés par les composantes principales du spectre. *Journées d'Études sur la Parole (JEP)*, pages 331–334.

[Seck et al., 1999] Seck, M., Bimbot, F., et Delyon, B. (1999). Comparaison de critères de segmentation par detection de rupture, sur un signal sonore. $17^{ème}$ *colloque GRETSI sur le Traitement du Signal et des Images*, 4:989–992.

[Seck et al., 2000] Seck, M., Blouet, R., et Bimbot, F. (2000). The IRISA/ELISA speaker detection and tracking systems in the NIST'99 evaluation campaign. *Digital Signal Processing*, 10(1-3):154–171. Special Issue: NIST 1999 Speaker Recognition Workshop.

[Shewart, 1931] Shewart, W. A. (1931). *Economic control of quality of manufactured product.* Prinston, d. van nostrand reinold édition.

[Shiryaev, 1995] Shiryaev, A. N. (1995). *Probability.* Graduate Texts in Mathematics. Springer. Second edition.

[Shorack et A., 1986] Shorack, G. R. et A., W. J. (1986). *Empirical processes with applications to statistics.* Wiley.

[Shwartz et Weiss, 1995] Shwartz, A. et Weiss, A. (1995). *Large deviations for performance analysis.* Chapman & Hall.

[Siegmund, 1985] Siegmund, D. (1985). *Sequential analysis - tests and confidence intervals.* Series in Statistics. Springer, New York.

[Sönmez et al., 1999] Sönmez, K., Heck, L., et Weintraub, M. (1999). Speaker tracking and detection with multiple speakers. *European Conference on Speech Communication and Technology, Eurospeech'99*, 5:2219–2222.

[Stephens, 1994] Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, 43(1):159–178.

[Sundberg, 1973] Sundberg, R. (1973). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statis.*, 1:49–58.

[Titterington et al., 1985] Titterington, D. M., Smith, A. F. M., et Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* John Wiley & Sons.

[van Laarhoven et Aarts, 1987] van Laarhoven, P. J. M. et Aarts, E. H. L. (1987). *Simulated annealing : theory and applications*. D. Reidel Publishing Co.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutionnal codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

[Wellish, 1991] Wellish, H. H. (1991). *Indexing from A to Z*. Bronx, New York, H. W. Wilson.

[Wu, 1983] Wu, J. C. F. (1983). On the convergence properties of the EM algorithm. *The Anals of Statistics*, 11(1):95–103.

[Yakowitz et Spragins, 1968] Yakowitz, S. D. et Spragins, J. D. (1968). On the identifiability of mixtures. *The Anals of Mathematical Statistics*, 39(1):209–214.

Résumé

Mots-clés : indexation, multimédia, signal sonore, mélange de gaussiennes, modèle de mélange, test statistique, détection de rupture, rapport de vraisemblance généralisé, divergence de Kullback, détection d'une classe de sons, suivi d'une classe de sons.

Dans le contexte actuel du besoin croissant d'outils de gestion de documents multimédia, cette thèse a pour objet de définir, étudier, mettre en œuvre et évaluer des algorithmes de description de documents sonores pour l'indexation. Ces travaux se concentrent sur les trois problèmes suivants : la détection de changements de plage sonore, la détection d'une classe de sons, et le suivi d'une classe de sons.

La détection de changements de plage sonore consiste d'abord à calculer, à chaque instant, un indice de rupture à partir d'une statistique de test. Ensuite, un critère de rupture est utilisé pour mettre en évidence les extrema dominants de l'indice, sans faire appel à un seuil de réglage auxiliaire. Comme alternative au Rapport de Vraisemblance Généralisé (RVG) habituellement utilisé pour le test d'une rupture, nous proposons une autre statistique, basée sur une estimation de la divergence de Kullback entre les distributions avant et après rupture (*le maximum de divergence de Kullback empirique généralisée*). Ces deux statistiques sont comparées, d'une part sur leurs lois asymptotiques en absence de rupture, et d'autre part sur une tâche expérimentale de détection de ruptures de type parole/musique.

Pour la détection et le suivi d'une classe de sons, la distribution des paramètres acoustiques du document est modélisée par un mélange des distributions cible et non-cible. Deux statistiques sont alors dérivées de ce modèle de mélange : l'estimateur du maximum de vraisemblance du poids de la distribution cible et le RVG du test de détection. Ces deux statistiques sont évaluées et comparées à la statistique usuelle du rapport de vraisemblance cible/non-cible, sur une tâche de suivi de parole dans une émission documentaire (corpus AIM de l'INA, projet DiVAN) et sur des tâches de détection et de suivi du locuteur dans une conversation téléphonique (corpus Switchboard, évaluations NIST au sein du consortium ELISA). Les résultats obtenus mettent en évidence des différences de performances de ces statistiques selon la tâche traitée, mais également selon la méthode d'estimation des distributions des classes de sons.

Abstract

Keywords : **indexing, multimedia, sound signal, gaussian mixture, mixture model, statistical test, detection of abrupt changes, generalized likelihood ratio, Kullback divergence, detection of a class of sounds, tracking of a class of sounds.**

In the current context of an increasing need for tools in multimedia document management, the purpose of this work is to define, investigate, implement and evaluate algorithms for the description and indexing of audio document. This work focuses on the three following tasks : the detection of homogeneous sound segments, the detection of a given class of sounds and the tracking of a given class of sounds.

The detection of homogenous sound segments consists firstly in calculating, at each instant, an abrupt change index based on a test statistic. Secondly, an abrupt change criterion is used to locate the dominating extrema of the index, without calling for any auxiliary threshold. As an alternative to the well-known Generalized Likelihood Ratio (GLR) for abrupt change detection, we propose an other statistic, based on an estimation of the Kullback divergence between the distributions before and after change (*the maximum of generalized empirical Kullback divergence*). These two statistics are compared, on the one hand as regards their asymptotical distributions in the absence of change, and on the other hand on an experimental task of speech/music abrupt change detection.

For the detection and tracking of a class of sounds, the distribution of the acoustical features extracted from the document is modeled by a mixture of the target and non-target distributions. Two statistics are then derived from this mixture model : the maximum likelihood estimator of the target distribution weight and the GLR of the corresponding detection test. These two statistics are evaluated and compared to the classical target/non-target likelihood ratio statistic, on a speech tracking task in a documentary (AIM corpus from INA, DiVAN Project), and on speaker detection and tracking tasks in conversational telephone speech (Switchboard corpus, NIST evaluations within the ELISA Consortium). The results show differences in terms of performance between the statistics, according to the task and to the method used for estimating the distributions.