Détection et Reconnaissance des Sons pour la Surveillance Médicale

Dan Istrate

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
| _ / _ / _ / _ / _ / _ / _ / _ / _ / _ /

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité : « Signal, Image, Parole, Télécoms »

préparée au laboratoire CLIPS - IMAG (Communication Langagière et Interaction Personne-Système)
dans le cadre de l'**École Doctorale « Électronique, Électrotechnique, Automatique, Télécommunications, Signal »**

présentée et soutenue publiquement

par
Dan Mircea ISTRATE
le 16 décembre 2003

Titre :

DÉTECTION ET RECONNAISSANCE DES SONS POUR LA SURVEILLANCE MÉDICALE

Directeur de thèse : Eric CASTELLI
Codirecteur de thèse : Laurent BESACIER

JURY

M.	James L. Crowley Professeur, HDR, INPG, Grenoble	Président
M.	Gaël Richard Maître de conférences, HDR, ENST, Paris	Rapporteur
M.	Michael Ansorge Privat Docent, Université de Neuchâtel	Rapporteur
M.	Eric Castelli Maître de conférences, HDR, INPG, Grenoble	Directeur de thèse
M.	Laurent Besacier Maître de conférences, Université de Grenoble 1(UJF)	Codirecteur de thèse
M.	Pierre Yves Coulon Professeur, HDR, INPG, Grenoble	Examinateur

...à mes chers parents pour toutes ces années passées loin d'eux

...à Daniela pour m'avoir soutenu et pour sa compréhension

Remerciements

Je commence par remercier mes deux directeurs de thèse Monsieur *Eric Castelli* et Monsieur *Laurent Besacier* pour avoir accepté d'encadrer cette thèse. Pour tous ses conseils et critiques sur le plan scientifique et aussi bien pour la relation humaine qu'il a su développer, je remercie chaleureusement Monsieur Eric Castelli. Je remercie aussi vivement Monsieur Laurent Besacier pour son intense participation à l'orientation de mes travaux de recherche, en tant que co-encadrant. Finalement, je voudrais remercier tous les deux pour toutes leurs critiques et conseils prodigués pendant la rédaction de ce manuscrit.

Mes vifs remerciements vont à Monsieur *Jim Crowley*, président de ce jury de thèse, pour l'intérêt qu'il porte à mes travaux de recherche. Je veux exprimer ma profonde reconnaissance à Monsieur *Gael Richard* et Monsieur *Michael Ansorge* pour avoir accepté d'être rapporteurs de cette thèse. Je remercie également Monsieur *Pierre Yves Coulon* d'avoir accepté être examinateur de cette thèse.

J'adresse mes remerciements à Monsieur *Jean François Sérignat*, responsable de l'équipe GEOD pour m'avoir accueilli dans son équipe. Je lui suis reconnaissant pour ses avis sur mes travaux de recherche, pour avoir lu mon manuscrit de thèse et pour ses commentaires utiles.

Je tiens à remercier Monsieur *Michel Vacher* avec qui j'ai travaillé sur le projet, pour les connaissances scientifiques et non scientifiques transmises pendant les nombreuses discussions que nous avons eu.

J'ai une mention spéciale pour Monsieur *Dominique Vaufreydaz*, pour ses conseils en informatique et pour sa bonne humeur. Je profite de cette occasion pour remercier l'ensemble des membres du laboratoire CLIPS et plus particulièrement les membres de l'équipe GEOD qui m'ont accepté et bien accueilli. Je remercie également le personnel administratif du laboratoire pour son efficacité.

Je remercie *mes parents* qui me sont très chers, pour avoir cru en moi, pour m'avoir soutenu et sans qui je n'aurai pas eu la possibilité d'effectuer cette thèse.

Je remercie *Daniela* pour sa présence, compréhension et bonne humeur. Je tiens à lui faire savoir combien sa présence a été précieuse.

J'ai une mention très spéciale pour tous mes amis : *Valeriu Vrabie, Quoc Cuong Nguyen* (mon collègue de bureau), *Daniel Moraru* et plus particulièrement pour *Adrian Staii* avec qui j'ai passé beaucoup de temps, qui m'a beaucoup aidé avec ses connaissances en anglais et de la vie. Je remercie aussi *Gilles Virone* pour le temps passé ensemble et les discussions plus ou moins liées au travail.

Un grand merci à tous
Dan Istrate

Résumé

Depuis quelques années se développe le concept général d'espace perceptif ou salle intelligente qui répond de diverses façons aux besoins, demandes, attentes des acteurs humains. Les espaces perceptifs traitent des signaux de parole, des signaux vidéo, les données de l'environnement, la localisation des personnes, le suivi et la reconnaissance des gestes, etc.

Ce travail de thèse se situe à la frontière entre les espaces perceptifs et la télémédecine qui a récemment évolué vers la télésurveillance des malades, le télédiagnostic, etc. La télésurveillance est notamment utilisée pour suivre l'évolution de personnes à risques (maladies chroniques ou personnes exposées à des situations critiques). Cela peut être à domicile (personnes âgées) ou dans un environnement professionnel dangereux.

L'analyse et l'extraction des informations du son est un aspect important des espaces perceptifs pour la télésurveillance médicale. Dans ce contexte, cette thèse analyse et propose des solutions aux problématiques spécifiques au traitement du son dans les espaces perceptifs plus particulièrement pour la télésurveillance médicale.

Parmi ces problématiques la classification automatique de sons de la vie courante a été très peu explorée jusqu'à aujourd'hui. Dans ce travail, un système d'analyse sonore en deux étapes est proposé pour éviter d'analyser un flux audio continu. Le rôle de la détection des événements sonores est d'extraire du bruit environnemental les signaux à identifier. Appliquée en même temps sur un ensemble de capteurs sonores répartis dans l'appartement, elle permet également une première localisation de la source sonore. Les algorithmes issus de l'état de l'art se montrant insuffisamment efficaces dans nos conditions, de nouveaux algorithmes mieux adaptés aux signaux impulsionnels, comme ceux utilisant la transformée en ondelettes sont proposés.

Pour la classification des sons proprement dite, l'utilisation de techniques issues de la reconnaissance automatique de la parole est d'abord envisagée. Ces techniques sont ensuite enrichies par l'ajout de paramètres acoustiques mieux adaptés, parmi lesquels ceux issus de la transformée en ondelettes et de la détection de signaux musicaux. Les performances de la classification sont aussi évaluées dans le bruit et une solution de pré-traitement est présentée.

Les problématiques liées au couplage entre la détection et la classification, ainsi que le problème de l'évaluation d'un tel système sont aussi abordées dans ce travail. En fin de manuscrit, l'évolution vers un système de reconnaissance de «sons clés», inspirée de la reconnaissance de mots clés en parole, est ébauchée.

Une implémentation en temps réel des algorithmes proposés a été réalisée pour l'application de télésurveillance médicale et est en cours de validation dans l'appartement test disponible pour le projet. Certains résultats expérimentaux présentés dans le document proviennent directement de cet appartement test.

Mots-clés : détection, classification, GMM, ondelettes, paramètres acoustiques, télémédecine, espaces perceptifs, traitement du signal.

Abstract

From few years, the general concept of perceptive spaces or smart rooms that answers in different way to the human actors needs, demands or expectations is in a continuous developing. The perceptive spaces deals with speech recognition, video signals, environmental data, persons localization, gesture following and recognition, etc.

The work presented in this thesis is set on the border of the perceptive spaces and telemedecine which has recently evolved to : telesurgey, medical telemonitoring, telediagnosis, etc. Telemonitoring as one of his branches, is used especially to follow the evolution of person with accident risk, that suffer of chronic diseases or persons exposed to critical situations. This can be applied not only at home for eardely but also in a dangerous professional environment.

The sound analysis and information extraction are important aspects of perceptives spaces for the medical telemonitoring. Thus, this thesis analyzes and proposes solutions to sound processing problems for the perceptives spaces, generaly and for medical telemonitoring, particulary.

From all the problems linked to perceptive spaces, the automatic classification of everyday life sounds was not explored too much until now. A two steps sound analysis system for avoiding the classification of a continuous audio flow is proposed in this work. The role of the sound event detection, the first step of the proposed system, is to extract the signal to be identified from the environmental noise. If the detection method is applied simultaneously to a sound sensors array that are distributed in an appartment, it allows also a first localization of the sound source. The state of the art algorithms are not efficient enough in our work conditions and thus, new algorithms, like those using the wavelet transform, better adapted to impulsive signals are proposed.

Concerning the sound classification itself, the second step of the proposed system, a first approach was the use of the automatic speech recognition techniques. These techniques are, then, improved by adding better adapted acoustical parameters among which those determined with wavelet transform and those used to detect the musical signals. The performances of the classification method are determined in a noisy environment and a preprocessing solution is presented too.

The problems concerning the coupling of detection and classification steps, as well as the system evaluation are also presented. In the last part of the thesis the evolution to a sound key recognition system is approached.

A real-time implementation of the proposed algorithms was realized for a medical telemonitoring application beeing in a validation process in the test appartment. Same experimental results obtained in the test appartment are presented in this work.

Key words : detection, classification, GMM, wavelets, acoustical parameters, telemedicine, perceptive spaces, signal processing

Sommaire

Table des figures

Liste des tableaux

Chapitre 1

Le contexte de la thèse

1.1 Introduction

Depuis quelques années se développe le concept général de «salles intelligentes» dans lequel il s'agit de concevoir des salles (salle de réunion, bureau d'étude, cuisine, chambre hospitalière, etc.) dotées de capteurs divers (microphones, caméras, détecteurs de présence infrarouge, micro-capteurs mobiles de signaux biologiques, etc.) gérés par un système informatique. Ces systèmes analysent les signaux soit en temps réel, soit après l'acquisition, et répondent de diverses façons (alarmes sonores, réponses vocales, changement des paramètres de l'environnement, etc.) aux besoins, demandes, attentes des acteurs humains. L'analyse temps réel concerne les systèmes qui doivent répondre aux besoins des utilisateurs ou qui doivent détecter des situations spéciales. L'analyse après l'acquisition concerne la segmentation des bases de données vidéo et audio pour faciliter la recherche d'informations.

Tous ces systèmes sont liés au concept «d'objets communicants». D'après le livre «Objets communicants» de Claude Kintzig et al. ([Kintzig et al., 2002]), un «objet communicant» est : «*un objet physique interagissant directement (c'est-à-dire par le biais de capteurs/actionneurs) ou par le biais de réseaux de communication de nature quelconque, avec son environnement physique, d'autres objets communicants et/ou des utilisateurs humains éventuels, doté au minimum de capacités de mémorisation numérique d'état, et, le cas échéant, de capacités de traitement numérique*». Les «objets communicants» deviennent présents partout, dans notre environnement, dans les réseaux, dans les systèmes de surveillance, etc.

L'évolution vers les «espaces perceptifs» (salles intelligentes, télésurveillance) implique l'apparition de nouvelles problématiques dans des domaines comme la reconnaissance automatique de la parole (RAP) et/ou du locuteur et celui des algorithmes de traitement du signal. Par exemple, pour la RAP, l'enregistrement du signal était autrefois déclenché par le locuteur en appuyant sur un bouton («push button»). Dans l'optique des espaces perceptifs, l'enregistrement du signal est continu et c'est le système qui segmente le signal. Une autre nouvelle problématique est celle de l'acquisition, non seulement en continu sur un canal mais sur plusieurs canaux en même temps, ce qui implique une grande quantité des données à traiter.

Mais dans les espaces perceptifs ne sont pas seulement traités les signaux de parole et/ou signaux vidéo mais aussi d'autres signaux comme ceux qui représentent les données de l'environnement (mesure de la quantité de lumière, de la température, etc.), la localisation des

personnes, le suivi et la reconnaissance des gestes, etc.

L'utilisation des microphones omnidirectionnels situés à des distances variables du locuteur impose le développement d'algorithmes de traitement du signal pour pallier les effets du bruit qui est plus important dans ce cas, pour éliminer l'écho qui est capté avec ce type de microphone, etc. En conclusion, comme le montre la figure 1.1, le passage vers les espaces perceptifs implique le traitement d'une grande quantité d'informations (nous pouvons citer la salle intelligente de NIST où sont enregistrés 1Go/min de données [Stanford et al., 2003]), sur plusieurs canaux en même temps et le traitement de signaux de qualité inférieure.

Flux du signal contrôlé	$\Longrightarrow$	*Flux continu*
Un seul canal	$\Longrightarrow$	*Multi-canaux*
Un seul type d'information	$\Longrightarrow$	*Plusieurs types de signaux (sonores, localisation, environnement,...)*
Bonne qualité du signal	$\Longrightarrow$	*Signaux bruités, amplitudes variables, bande de fréquence limitée*

Fig. 1.1: Évolution du domaine de la reconnaissance de la parole/locuteur vers les espaces perceptifs

Le sujet de cette thèse, la détection et la classification des sons pour la surveillance médicale, touche en même temps au domaine de la télémédecine et des problématiques des espaces perceptifs. Dans les sections qui suivent, nous présentons ces domaines avec leurs principaux axes de recherche, le projet auquel était rattachée cette thèse, ainsi que la position et les problématiques de ce travail. Nous finirons ce premier chapitre avec une description de l'organisation du document.

1.1.1 Espaces perceptifs et salles intelligentes

Une application prédominante dans les espaces perceptifs semble être la mise en oeuvre de ressources facilitant les réunions de travail. Les réunions entre les individus font partie intégrante de la vie courante de tous les groupes de travail. Cependant, la participation sur place aux réunions de travail est souvent rendue difficile pour des raisons de temps requis par les déplacements, de surcharge des agendas ou pour d'autres contraintes.

Les téléconférences peuvent résoudre ce problème en évitant les déplacements tout en permettant l'interactivité entre participants. Par ailleurs, les enregistrements sonores et video permettent aux personnes ne pouvant se libérer de s'informer, tout en servant à établir des archives.

La recherche d'informations dans les grandes bases de données d'enregistrements de réunions est difficile et impose l'utilisation d'algorithmes efficaces de segmentation et d'étiquetage de données sonores et vidéos. D'autre part, dans le cadre des réunions, un environnement capable de s'adapter automatiquement aux besoins des utilisateurs ou par commande est de plus en plus demandé.

L'apparition de ces nouvelles problématiques ajoutée au développement des ordinateurs et des microcontrôleurs font qu'aujourd'hui dans le domaine des espaces perceptifs il y a de plus en plus de laboratoires de recherche impliqués.

La première problématique étudiée est celle de l'*enregistrement sonore et vidéo des réunions* avec une bonne qualité, comme par exemple dans le projet de Microsoft [Cutler, 2003]. L'enregistrement se fait dans un contexte où plusieurs personnes parlent en même temps ou à tour de rôle. Ces contraintes imposent la présence de plusieurs caméras et microphones. Des systèmes de suivi de la personne existent pour les caméras vidéos ; pour les microphones, soit on enregistre avec un microphone par personne, soit on utilise une antenne acoustique avec suivi automatique du locuteur comme dans le cas du projet du laboratoire NIST[1]. Leur antenne acoustique est constituée de 64 microphones, traités en temps réel pour suivre le locuteur et par conséquent réduire le bruit [Stanford et al., 2003].

Pour la salle intelligente de Microsoft [Cutler, 2003], l'indexation de la personne qui parle pendant l'enregistrement, est faite en utilisant le son (la localisation acoustique) et l'image (détection et suivi de personne). Le suivi de la personne en temps réel est utile pour réduire la taille de l'enregistrement et pour faciliter la tâche des algorithmes de post-traitement.

La standardisation des bases de données, comme problématique de l'enregistrement des données issues d'espaces perceptifs, est un des axes de recherche de l'institut NIST. La présence d'une multitude de signaux de différents types (sonores, vidéo et d'autres capteurs d'environnement) rend difficile la classification et la fusion de ces données.

Les salles intelligentes doivent également être capables d'inclure différentes technologies : des ordinateurs portables, des badges intelligents de localisation, des réseaux sans fil avec le problème de la sécurité des données, de nouvelles interfaces homme-machine, etc.

La reconnaissance de la parole et/ou celle du locuteur font partie des principaux traitements du son. Le bruit, la réverbération et la présence de la parole de plusieurs locuteurs réduisent beaucoup les taux de reconnaissance des algorithmes actuels [Renals and Ellis, 2003]. Pour la salle intelligente du laboratoire IDIAP[2] [McCowan et al., 2003], l'axe principal de recherche est celui de la transcription textuelle de la parole enregistrée. Le but de la transcription en texte est de faciliter la recherche d'informations ultérieure sur une base de documents multimodaux (son, image, texte). Cette même problématique est étudiée aussi par ICSI[3] pour l'application de leur salle intelligente [Morgan et al., 2003]. La segmentation et le suivi des locuteurs (aidé éventuellement par le traitement de l'image) sont nécessaires pour la segmentation de l'enregistrement en tours de parole, soit pour pouvoir facilement trouver les interventions d'une personne particulière, soit pour réaliser des statistiques sur les temps de parole des participants. L'identification de la langue et de l'accent dans le cadre du projet d'espace perceptif du laboratoire ISL[4] est également abordée. La suppression des hésitations et/ou des répétitions par un système linguistique est aussi étudiée dans le même projet [Waibel et al., 2003].

Un autre axe de recherche abordé dans les espaces perceptifs est celui de la *reconnaissance des gestes* en vue de la compréhension des interactions humaines. L'institut de recherche IDIAP[2] s'occupe de l'analyse des gestes dans la surveillance vidéo à l'intérieur des bureaux. L'étude du comportement et des gestes des individus peut se faire par deux méthodes : soit par une reconnaissance des gestes de chaque individu suivi d'une fusion de haut niveau pour toutes les personnes ; soit par une reconnaissance des gestes de chaque individu en tenant compte des contraintes imposées par les gestes des autres individus.

[1] National Institute of Standards and Technology, USA
[2] Institut Dalle Molle pour l'Intelligence Artificielle perceptive, Martigny, Suisse
[3] International Computer Science Institute, Berkeley, USA
[4] Interactive Systems Laboratories, Karlsruhe, Allemagne

Une possibilité, pour faciliter la tâche du système de post-traitement des enregistrements des réunions est d'utiliser les notes prises par les participants comme des *index* naturels. Le laboratoire FXPAL[5] a imaginé et développé une application client serveur, NoteLook. NoteLook permet à l'utilisateur de prendre des notes manuscrites et d'y incorporer de manière interactive des images de la caméra sélectionnée et du document courant projeté sur le tableau [Chiu et al., 2001]. L'application cliente tourne sur des tablettes PC dispersées sur les tables.

Le laboratoire MIT[6] étudie l'intégration de capteurs pour détecter les mouvements dans la pièce, les niveaux de lumière et de bruit, la pression exercée sur des surfaces telles que le plancher et les meubles [Wren et al., 1997]. Ils ont crée Look To Talk, un système qui utilise le regard comme une interface pour activer la reconnaissance vocale automatiquement.

Plus localement, le laboratoire CLIPS-IMAG en collaboration avec le CNRS, le CEA, France Telecom, ST Microelectronics et IDEAS LAB participe au projet COUCOU (Conception participative Orientée Usage de services de Communication et d'objets Ubiquistes) qui fait partie lui aussi des projets liés à la problématique des espaces perceptifs. Plus précisément, son but est d'implémenter deux salles intelligentes de réunions avec les outils pour prendre des notes et pour le post-traitement des données enregistrées.

Comme post-traitement des bases de données multimodales, *la segmentation des enregistrements sonores et vidéo* est un autre axe de recherche [Sundaram and Chang, 2000]. Cette segmentation est nécessaire pour les besoins de moteurs de recherche dans les bases de données et pour un bon archivage des réunions. La segmentation peut se faire pour le signal audio en utilisant divers critères. Par exemple, une première segmentation est faite en fonction du type de signal : parole, musique ou bruit suivi, éventuellement, d'une nouvelle segmentation de chaque classe. Le signal de parole peut être divisé soit en locuteurs soit en mots, propositions, phrases. Le signal de musique peut être utilisé pour la reconnaissance du nom de la mélodie ou peut être classifié en fonction du type de musique. Au niveau du signal vidéo, la segmentation se fait en plans vidéos. Aujourd'hui, la tendance est de faire fusionner la segmentation sonore avec celle provenant de la vidéo pour obtenir de meilleurs résultats.

1.1.2 La télémédecine

La télémédecine est un concept général qui couvre différentes applications en rapport avec la santé. Selon l'**O**rganisation **M**ondiale de la **S**anté (OMS), la télémédecine couvre l'utilisation d'informations et de techniques de communication dans les systèmes de santé pour des soins donnés directement ou indirectement [Kornblum et al., 2001].

La télémédecine est née dans les années 1970 dans l'Amérique rurale (Nebraska, Texas et Géorgie) et dans le nord de la Norvège. Elle a également été expérimentée très tôt au Canada. L'objectif était alors d'assurer à distance des soins de qualité comparables à ceux pratiqués dans les centres urbains. Plus récemment, la télémédecine a connu des développements visant à résoudre des problèmes d'organisation sanitaire comme : l'accès des personnes habitant dans des endroits isolés aux consultations des spécialistes, réduction des coût de la sécurité sociale par la télésurveillance au domicile des malades chroniques à la place des longues périodes d'hospitalisation [Fieschi, 2000].

Une définition possible de la *télémédecine* est l'utilisation des nouvelles techniques de l'in-

[5] Fuji-Xerox Palo Alto Laboratory, USA
[6] Massachusetts Institute of Technology, USA

formation, de la communication pour des applications médicales. Cela implique l'acquisition des données leur stockage ou leur transmission en temps réel ou non, par réseaux avec ou sans fil, la fusion et le traitement automatique des informations. Comme données acquises nous pouvons avoir : des grandeurs purement médicales (pression artérielle, oxymétrie, etc.), des informations de position de la personne, *des informations extraites du son* ou de l'image, des paramètres de l'environnement (température, pression, quantité de lumière, etc.). Mais la télémédecine implique aussi la transmission des informations dans le sens centrale vers le patient.

Autrement dit, la télémédecine peut être définie comme la consultation et la surveillance des patients en utilisant des systèmes qui donnent un accès rapide et facile à l'expert médical et au patient quelle que soit la localisation du patient et de l'expert [Chevrolet et al., 2002].

Jusqu'aux années 2000 (conformément à [Kornblum et al., 2001]) la télémédecine était centrée autour de la transmission de l'image. Aujourd'hui, la télémédecine a évolué rapidement. Ceci est illustré par le nombre croissant de projets, l'apparition de formations de type DEA dédiées à la télémédecine dans le cadre des universités et des Grandes Écoles (DEA *Méthodologie des réseaux en médecine et Traitement des images en télémédecine* de l'Université Paris 6), l'existence d'une société Internationale de télémédecine - ISfT (International Society for Telemedicine).

La confidentialité des données et la sûreté des algorithmes du traitement automatique sont aussi des aspects importants pour un système de télémédecine parce que le degré de confidentialité de ces données est important pour protéger les libertés individuelles. La fiabilité du cryptage des données lors de la transmission par réseaux, est très importante. Ce problème est résolu par les algorithmes développés pour les transferts bancaires sur réseaux. Si dans le cas des signaux sortis de capteurs purement médicaux, une analyse fiable est déjà mise au point, l'extraction des informations issues des capteurs sonores est encore loin de pouvoir respecter les exigences médicales. Les capteurs sont considérés comme intelligents si ils incorporent les pré-traitements des signaux. Une problématique importante des capteurs médicaux est celle du bruit de mesure. L'information utile des capteurs sonores (le contenu linguistique ou le type de son) est obtenue après des traitement assez compliqués sur le signal des microphones. Une fusion de données de plusieurs types de capteurs peut alors s'avérer fiable par la redondance des informations : par exemple la chute de la personne identifiée par un capteur de chute est confirmée par la reconnaissance du son spécifique de chute.

Parmi les applications qui ont déjà fait leurs preuves et qui sont promises à des développements intéressants nous distinguons :

- ✓ **Télésurveillance :** La surveillance de paramètres physiologiques peut être faite en temps réel ou de manière différée si les données sont mémorisées. La surveillance est utilisée pour suivre l'évolution de personnes à risques (maladies chroniques ou personnes exposées à des situations critiques). Cela peut être à domicile (personnes âgées) ou dans un environnement professionnel dangereux.
- ✓ **Télédiagnostic :** Il s'agit d'applications, telles que télé-électrocardiogrammes, télé-dermatologie ou télé-endoscopie, pour lesquelles un examen médical est fait par ou avec la participation d'un médecin, qui se trouve en un autre lieu.
- ✓ **Téléconsultation :** Elle est très utile pour obtenir un deuxième avis d'un autre médecin pour des cas de pathologies complexes qui apparaissent rarement ou qui présentent des risques élevés. Elle peut aussi intervenir pendant une opération.

- ✓ **Téléréunion de travail :** Même sur de faibles distances, par exemple dans un hôpital, la télémédecine permet d'échanger efficacement des informations entre les différents donneurs de soins.
- ✓ **Télétriage (Pre-gatekeeping) :** Grâce à des centres d'appel médicaux, il est possible d'effectuer de manière professionnelle à distance l'analyse de certains problèmes médicaux et d'acheminer rapidement les patients à l'endroit le plus approprié compte tenu des symptômes observés.
- ✓ **Téléservices pour cas d'urgence :** Dans des situations d'urgence médicale, l'importance des centraux téléphoniques de secours peut être renforcée en ayant accès à des experts, afin d'aider à prendre des décisions lors de l'évaluation initiale des données médicales transmises concernant des cas graves.
- ✓ **Télé-opération :** Normalement, les hôpitaux et les chirurgiens sont disponibles, une téléopération n'est essentiellement intéressante que pour intervenir chirurgicalement à distance dans une région isolée, avec des instruments qui peuvent être télécommandés (télérobotique).
- ✓ **Télé-éducation et Télé-formation (eLearning) :** Les professionnels de la santé qui sont déjà formés peuvent continuer leur formation par ce moyen qui est fortement lié aux technologies de l'information et de la communication.

L'analyse des différentes applications conduit à définir trois catégories principales de systèmes de télémédecine :

1. Transfert et gestion d'informations médicales entre les acteurs du domaine de la santé, via le réseau public de communications (télématique de la santé)
2. Télésurveillance en temps réel ou en différé de paramètres physiques, physiologiques et pathologiques
3. Contrôle à distance de procédures médicales

Une des problématiques de la télémédecine est la *compression* et la *transmission des images médicales* par le réseau Internet. La résolution des images médicales est grande et importante ce qui augmente la taille des fichiers à transmettre et impose des pertes réduites de qualité à la compression [VBCH Project, 2002], [Lee et al., 1994]. Les réunions de travail réalisées par téléconférence constituent une autre application de la transmission des images et du son par réseau [LITMED Project, 2002].

La *télésurveillance médicale* à l'hôpital, à domicile ou au travail est un axe de recherche pour la télémédecine. A domicile la surveillance continue ou périodique des grandeurs médicales nécessite des capteurs spécialisés de dimensions réduites, robustes, autonome (avec alimentation mixte : piles et secteur), avec transmission des informations au centre de surveillance. Une autre application à domicile est la distribution surveillée et contrôlée par réseau des pilules pour les personnes âgés [Olsen et al., 2003]. A l'hôpital la recherche est focalisée sur la surveillance des grandeurs médicales non classiques comme le mouvement des patients souffrant de la maladie d'Alzheimer [Projet TIISSAD, 2001], ou le son. La surveillance au travail nécessite des capteurs de taille réduite, à faible consommation électrique, robustes et dotés d'une transmission sans fil des informations. Un exemple est la surveillance du rythme cardiaque d'un athlète en mouvement [Vassiliadis, 2003].

Un autre axe de recherche de la télémédecine concerne les *interfaces homme-machine* des appareils médicaux de chirurgie [Dubois, 2001]. La robustesse de l'interface est devenue très importante à cause du nombre croissant des appareils et des informations mises à disposition des médecins.

La présence d'un nombre important de projets (voir annexe B de la page 153) et de laboratoires dans le domaine de la télémédecine démontre que le domaine est en plein développement [Lau et al., 2002]. D'autre part, la télésurveillance médicale est peut être une piste pour réduire l'isolement de certaines personnes âgées, particulièrement exposées en période de canicule comme ce fut le cas en Août 2003 en France [Prieur, 2003].

Jusqu'à présent, on observe l'absence de l'utilisation du son dans le cadre de la télésurveillance médicale. Dans ce contexte l'extraction d'informations lors d'une surveillance sonore pourrait donner des résultats intéressants par identification des sons suspects de chute ou en aidant l'identification des pathologies comme les troubles urinaires nocturnes.

1.2 Le projet RESIDE - HIS

Le projet **RE**connaissance de **SI**tuations de **DÉ**tresse en **H**abitat **I**ntelligent **S**anté (RESIDE-HIS) est labelisé IMAG[7]. C'est une collaboration entre l'équipe GEOD[8] du laboratoire CLIPS[9]-IMAG et l'équipe AFIRM[10] du laboratoire TIMC[11]-IMAG [Castelli et al., 2002] .

Nos travaux de recherche, se sont déroulés dans le cadre des activités de l'équipe GEOD pour développer ce projet.

L'objectif du projet est la conception, la mise au point et l'expérimentation d'un dispositif de télémédecine s'appuyant sur l'utilisation de capteurs de déambulation et d'activité, et l'utilisation d'algorithmes de classification automatique des sons et de la parole pour détecter, identifier et interpréter des situations de détresse [Projet RESIDE-HIS, 2000]. L'originalité de l'application se situe dans le remplacement de la surveillance vidéo (mal acceptée par les patients) par l'utilisation non conventionnelle de signaux de parole (tels que des gémissements, cris, des appels au secours) et des sons de la vie courante qui viendront compléter les informations données par les capteurs d'activité. Le dispositif prend place au sein d'un habitat intelligent et a pour but, à terme, de contribuer au maintien de patients à leur domicile en transmettant les alarmes vers les centres de médico-surveillance.

Un local du laboratoire TIMC a été entièrement équipé pour en faire un Habitat Intelligent pour la Santé pilote, à des fins d'expérimentation et de simulation. Cette réalisation constitue un prototype d'appartement de type T1 (environ 30 m2), comprenant les zones d'habitat classiques que sont la chambre, le séjour, la cuisine, les toilettes, la douche et un couloir [Rialle et al., 1999]. Une zone technique attenante à l'appartement a été ajoutée afin de recevoir le système informatique d'expérimentation du projet. Le plan de l'HIS est visible sur la figure 1.2 illustrant la disposition des capteurs [Virone et al., 2002].

Le maintien au domicile de personnes dépendantes suppose la détection et l'analyse de situations de détresse. Les capteurs utilisés sont des capteurs d'activité physiologique et des

[7] Institut d'Informatique et de Mathématiques Appliqués de Grenoble
[8] Groupe d'Etude sur l'Oral et le Dialogue
[9] Communication Langagière et Interaction Personne-Système
[10] Acquisition, Fusion d'Informations et Réseaux pour la Médecine
[11] Techniques de l'Imagerie, de la Modélisation et de la Cognition

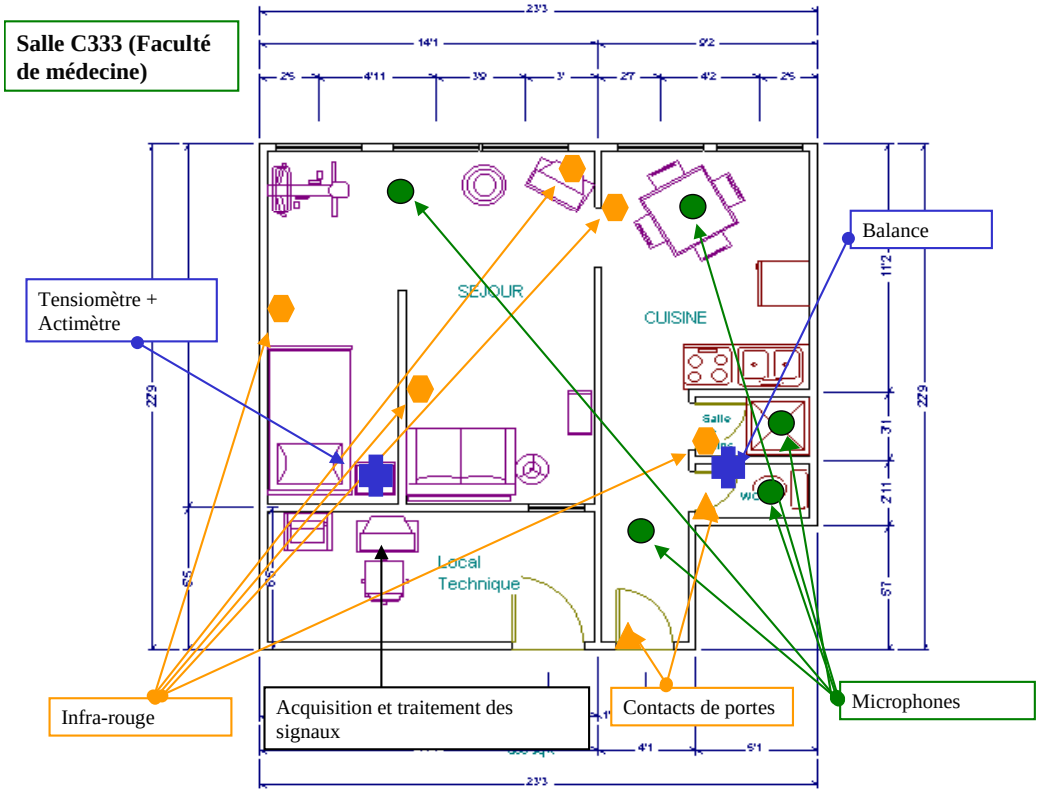


Fig. 1.2: Architecture de l'appartement et position des capteurs

capteurs sonores. Au niveau de l'analyse sonore, en plus de la localisation de la pièce où se trouve la personne, le but est de reconnaître parmi les sons de la vie courante, les éventuels appels au secours ou d'éventuelles alarmes audio (chute, bris de verre, etc.). Pour des contraintes liées à la préservation de la sphère privée de la personne le son est acquis et analysé en temps réel sans enregistrement continu et sans un système de reconnaissance continue de la parole. L'application médicale impose, par ailleurs, des contraintes de fiabilité du système.

L'ensemble des microphones constitue ainsi un capteur intelligent qui communique avec un ordinateur Maître. Celui-ci surveille les capteurs physiologiques et le capteur sonore et par une fusion de données des informations redondantes ou non, identifie une situation de détresse. Par exemple, la position de la personne est donnée par les capteurs de position infrarouges et par le capteur sonore.

1.3 Espace perceptif lié à la télésurveillance médicale

Un espace perceptif est un environnement capable de percevoir la présence et des commandes des personnes. Les salles intelligentes et les espaces liés à la télémédecine entrent dans cette classification.

Les salles intelligentes sont capables d'une part de recevoir des informations par l'intermédiaire de capteurs sonores, vidéos et environnementaux et d'autre part de modifier elles-même l'environnement par des actionneurs. Les signaux de différents capteurs sont traités en vue de

l'extraction de l'information intéressante (le texte de la parole, par exemple, pour un capteur sonore).

La télémédecine utilise les mêmes outils dans un but médical, soit pour émettre un diagnostic à distance, soit pour guider des opérations chirurgicales, soit pour surveiller un malade ou une personne à risque.

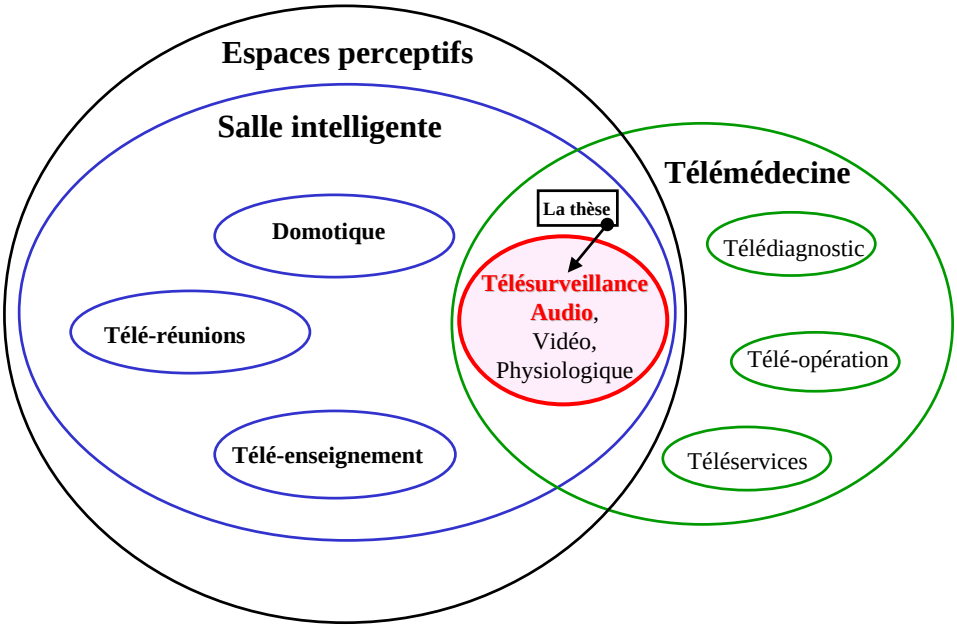


Fig. 1.3: La position de l'espace perceptif étudié

Du point de vue de la télémédecine, le système présenté est une partie d'une application de télésurveillance. Les applications actuelles de télésurveillance médicale des personnes âgées ou des malades en convalescence utilisent des capteurs vidéos. Nos recherches se sont portées sur l'utilisation de capteurs sonores comme source d'information à la place de l'utilisation d'une caméra vidéo. En effet, une caméra est moins bien acceptée par les patients car elle affecte davantage la préservation de la sphère privée. La source d'information des capteurs sonores peut être :

- une source directe :
  - reconnaissance d'un appel de détresse
  - détection d'un son suspect (chute d'objet, cris)
  - longue absence d'activité sonore pendant la journée
- une source indirecte :
  - détection d'une succession de sons de la vie courante qui par analyse sur une longue durée peuvent être un symptôme de pathologie comme les troubles urinaires nocturnes.

La détection d'une longue absence d'activité sonore en fonction de la période de la journée peut être un indicateur d'une situation de détresse grave. En France, nous connaissons quelques cas de personnes âgées décédées et découvertes bien plus tard par des voisins. Lors du journal

de 20.00h de TF1 (01.10.2003) une statistique des accidents domestiques a été presentée : parmi les 11000 accidents mortels des personnes âgées sur l'année 2003, 66% représentent des chutes.

Mettre en place une télésurveillance sonore, en utilisant aussi des capteurs physiologiques, est une opération qui combine les spécificités tant du domaine des salles intelligentes que du domaine de la télémédecine.

Ainsi, nos travaux de recherche se situent au niveau de la partie sensorielle des espaces perceptifs et plus particulièrement dans le domaine sonore. Entre la multitude des traitements de sons possibles, nous avons choisi la détection des sons dans le bruit et la classification des sons de la vie courante parce que les informations issues de ces traitements sont complémentaires à celles des capteurs médicaux pour l'application de télésurveillance médicale. Ces traitements peuvent être réalisés en temps réel sur plusieurs canaux pour un système interactif ou en temps différé pour l'archivage des enregistrements.

1.4 Objectifs et problématiques de la thèse

Cette thèse se situe dans le cadre des problématiques et des objectifs des espaces perceptifs liés à la télémédecine, elle a comme objectif principal l'étude et la validation d'un système multi-canaux d'identification et de classification de sons de la vie courante. Les catégories de sons étudiés devront pouvoir mettre en valeur l'état normal ou l'état de détresse du patient sous surveillance. L'étude de la fiabilité du système dans des conditions réelles de bruit environnemental, fait aussi partie des travaux de cette thèse. La complexité du système proposé sera limitée par la contrainte du traitement en temps réel.

Parmi les problématiques des espaces perceptifs et des applications de télémédecine, celles qui ont été étudiés dans cette thèse sont :

- La qualité du signal (la présence du bruit, larges gammes dynamiques, bandes de fréquence limitées, etc.)
- Le volume de l'information (5 canaux sonores acquis simultanément)
- Le compromis entre la complexité des algorithmes et les performances du système utilisé
- La réalisation d'un corpus des sons de la vie courante
- L'établissement d'une méthodologie d'évaluation d'un tel système
- L'adaptation des techniques de reconnaissance de la parole/du locuteur à la classification des sons de la vie courante

La qualité du signal influence les performances des systèmes de reconnaissance et de traitement du signal. Tenant compte des difficultés liées à la qualité du signal, le travail de cette thèse s'est focalisé sur les signaux limités en fréquence (bande de fréquence de 8 kHz limitée par la carte d'acquisition utilisée et le nombre des canaux nécessaires), bruités et ayant une grande dynamique. Cette bande passante est classiquement utilisé dans la reconnaissance de la parole. Ces signaux pourront ainsi être utilisés pour la reconnaissance des expressions de détresse.

L'acquisition et le traitement simultané des signaux provenant de plusieurs canaux sonores fait partie des nouvelles problématiques des espaces perceptifs. La présence d'un grand volume d'information à traiter instantanément impose l'utilisation d'ordinateurs puissants, d'algorithmes rapides et l'élimination de l'information redondante. Une solution à ce problème est

de rechercher sur tous les canaux sonores les signaux importants pour l'application et d'analyser seulement ces signaux, en réalisant ainsi un compromis entre la puissance actuelle des ordinateurs et les nécessités de temps réel des applications.

Une autre problématique importante est celle du *compromis entre la complexité des algorithmes et les performances* du système en terme de taux d'erreurs. En effet, par les raisons de confidentialité enoncées plus haut, le système de détection et de classification du signal fonctionne en temps réel. Les algorithmes doivent donc être de faible complexité, *mais* tout en garantissant des performances suffisantes pour une application aussi critique que la télémédecine.

Pour pouvoir tester, améliorer et valider les algorithmes proposés dans cette thèse nous avons besoin d'un corpus de sons spécifique aux applications médicales. Au début de notre travail un tel corpus n'existait pas, alors nous avons dû l'enregistrer. *La méthodologie de l'enregistrement* d'un tel corpus est une autre problématique. Pour un corpus de sons, la qualité des signaux est imposée et il est difficile de produire certains sons de détresse imposés par l'application de télésurveillance médicale.

La *méthodologie d'évaluation des performances* du système global et de ses parties est aussi importante pour garantir un niveau de robustesse à des applications critiques, comme c'est le cas pour les applications en télémédecine. Les méthodologies existant dans le domaine de la reconnaissance de la parole ou du locuteur devront être adaptées à notre problème.

Une dernière problématique abordée dans cette thèse est *l'adaptation des techniques issues de la reconnaissance de la parole ou du locuteur à la classification des sons de la vie courante*. Les caractéristiques fréquentielles de la parole et celles des sons de la vie courante sont différentes ce qui nécessite de trouver des paramètres acoustiques adaptés pour enrichir les méthodes classiquement utilisées en parole.

1.5 Organisation du document

Au vu des objectifs et des problématiques de la section précédente, nous proposons l'organisation suivante pour ce manuscrit de thèse.

Après une courte introduction, ce document continue avec une description du corpus de sons (figure 1.4). Ce corpus de sons, décrit dans le chapitre 2, a été réalisé avec des signaux enregistrés dans le studio du laboratoire et avec des enregistrements issus d'autres bases de sons. La méthodologie de réalisation des bases de test pour valider la détection, la classification et, respectivement, le couplage détection/classification est présentée aussi dans ce chapitre.

Le chapitre 3 aborde la détection des événements sonores dans le bruit. Nous commençons avec la description de trois algorithmes issus de l'état de l'art et leurs performances en conditions réelles. La méthodologie d'évaluation des algorithmes sur la base de tests est présentée. Les performances des algorithmes classiques s'avérant inacceptables, trois nouveaux algorithmes de détection sont proposés dans la section suivante.

Le chapitre 4 commence par une présentation des techniques existantes pour la classification des sons. Nous continuons avec une étude statistique sur la pertinence des paramètres acoustiques existants. Une proposition de nouveaux paramètres mieux adaptés aux types de sons étudiés est faite. Parmi ces nouveaux paramètres, des paramètres issus de la transformée en ondelettes sont proposés. A la fin de ce chapitre, l'influence du bruit sur le processus de classification des sons de la vie courante est évaluée. Des paramètres «résistants aux bruits»

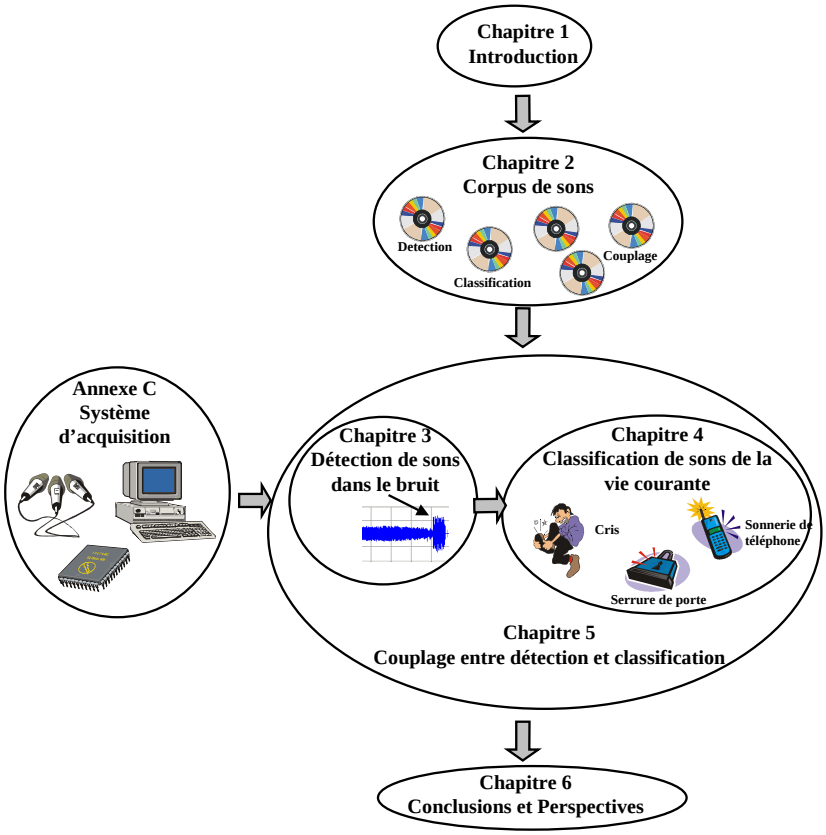


Fig. 1.4: Plan de la thèse

(ceux issus de la transformée en ondelettes) ainsi que le débruitage des signaux avant la phase de classification (par un seuillage de la transformée en ondelettes) sont proposés.

Le couplage détection/classification est abordé dans le chapitre 5 avec une solution d'amélioration de résultats.

Le document se termine par les conclusions dans le chapitre 6 et les annexes. Dans les annexes sont présentées les publications de l'auteur, quelques rappels mathématiques et la description technique du système d'acquisition sonore multivoies.

Bibliographie

[Castelli et al., 2002] Castelli, E., Serignat, J., and Rialle, V. (2002). Rapport final du projet RESIDE-HIS (Reconnaissance de situations de détresse en Habitat Intelligent Santé). Technical report, CLIPS et TIMC.

[Chevrolet et al., 2002] Chevrolet, J. C., Denz, M., Merminod, B., Osswald, S., and Roulet, M. (2002). Télémédecine CH. Technical report, Académie Suisse des Sciences Médicales.

[Chiu et al., 2001] Chiu, P., Boreczky, J., Girgensohn, A., and Kimber, D. (2001). Liteminutes : An Internet-Based system for multimedia meeting minutes. *World Wide Web Conference*, pages 140–149, Hong-Kong.

[Cutler, 2003] Cutler, R. (2003). The distributed meetings system. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 756–759, Hong-Kong.

[Dubois, 2001] Dubois, E. (2001). *Chirurgie augmentée, un cas de réalité augmentée ; Conception et réalisation centrées sur l'utilisateur*. PhD thesis, Université Joseph Fourier, Grenoble.

[Fieschi, 2000] Fieschi, M. (2000). Les médecins aussi tissent leur toile. *Supplément La Recherche*, pages 15–19.

[Kintzig et al., 2002] Kintzig, C., Poulain, G., Privat, G., and Favennec, P. N. (2002). *Objets communicants*. ISBN 2-7462-0475-4. LAVOISIER Hermès Science Publications, Paris.

[Kornblum et al., 2001] Kornblum, C., Sibony, O., Rol, A. L., Strauss, A., Champetier, D., Lavictoire, M., and Berry, M. (2001). Télémédecine & industrialisation. Technical report, Ministère de l'Emploi et de la Solidarité.

[Lau et al., 2002] Lau, C., Churchill, R. S., Kim, J., et al. (2002). Asynchronous web-based patient-centered home telemedicine system. *IEEE Transactions on Biomedical Engineering*, 49(12) :1452–1459.

[Lee et al., 1994] Lee, W., Kim, Y., Gove, R. J., and Read, C. J. (1994). MediaStation 5000 : Integrating video and audio. *IEEE Multimedia*, 1(2) :50–61.

[LITMED Project, 2002] LITMED Project (2002). Litmed II Project. http ://www.litmed.net/.

[McCowan et al., 2003] McCowan, I., Bengio, S., Gatica-Perez, D., et al. (2003). Modeling human interaction in meetings. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 748–751, Hong-Kong.

[Morgan et al., 2003] Morgan, N., Baron, D., Bhagat, S., et al. (2003). Meetings about meetings : Research at ICSI on speech in multiparty conversations. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 740–743, Hong-Kong.

[Olsen et al., 2003] Olsen, B. I., Eggen, A. E., Bellika, J. G., et al. (2003). An electronic health record-based system for automatic monitoring and control of medication. *International Conference on Information Communication Technologies in Health, Samos, Greece*, pages 147–151.

[Prieur, 2003] Prieur, C. (2003). L'été le plus meurtrier en france. *Le Monde, 9 Septembre*.

[Projet RESIDE-HIS, 2000] Projet RESIDE-HIS (2000). Projet RESIDE-HIS. http ://www-clips.imag.fr/tech-adm/perso/michel.vacher/HIS/SITE_WEB/index.htm.

[Projet TIISSAD, 2001] Projet TIISSAD (2001). TIISSAD project. http ://www.loria.fr-/projets/TIISSAD/.

[Renals and Ellis, 2003] Renals, S. and Ellis, D. (2003). Audio information access from meeting rooms. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 744–747, Hong-Kong.

[Rialle et al., 1999] Rialle, V., Lauvernay, N., Franco, A., Piquard, J. F., and Couturier, P. (1999). A smart room for hospitalized elderly people : Essay of modelling and first steps of an experiment. *Technology and Health Care*, 7 :343–357.

[Stanford et al., 2003] Stanford, V., Garofolo, J., Galibert, O., Michel, M., and Laprun, C. (2003). The NIST smart space and meeting room projects : Signals, acquisition, annotation and metrics. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 736–739, Hong-Kong.

[Sundaram and Chang, 2000] Sundaram, H. and Chang, S. F. (2000). Audio scene segmentation using multiple features, models and time scales. *ICASSP, Istanbul, République Turque*.

[Vassiliadis, 2003] Vassiliadis, K. (2003). DROMEAS - a wearable platform for the monitoring of health condition and sport performance of athletes and the real-time prevention of sport injuries. *International Conference on Information Communication Technologies in Health, Samos, Greece*, pages 136–140.

[VBCH Project, 2002] VBCH Project (2002). Van Buren County Hospital project. http ://showcase.netins.net/web/forhealth/telemed.htm.

[Virone et al., 2002] Virone, G., Noury, N., and Demongeot, J. (2002). A system for automatic measurement of circadian activity in telemedicine. *IEEE Transactions on Biomedical Engineering*, 49(12) :1463–1469.

[Waibel et al., 2003] Waibel, A., Schultz, T., Bett, M., et al. (2003). Smart : The smart meeting room task at ISL. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 752–755, Hong-Kong.

[Wren et al., 1997] Wren, C. R., Sparacino, F., et al. (1997). Perceptive spaces for performance and entertainment : Untethered interaction using computer vision and audition. *Applied Artificial Intelligence (AAI) Journal*, pages 267–284.

Chapitre 2

Corpus de sons de la vie courante

2.1 Introduction

Le but étant d'obtenir un système d'extraction d'informations sonores pour une application de télésurveillance médicale, le premier pas consiste à obtenir un corpus de sons. Le système d'extraction sonore étant divisé en deux parties, le corpus sonore servira autant pour la détection dans le bruit des événements sonores que pour la classification des sons de la vie courante. Cette division du système, permet une réduction de la quantité d'informations sonores par application de l'algorithme de détection (5 canaux sonores acquis en même temps à une fréquence d'échantillonnage de 16 kHz $\Rightarrow$ 80000 échantillons/s à traiter). La classification des sons intervient seulement sur les segments sonores détectés (figure 2.1). En fonction de la classe de sons identifiée le système enverra ou non une alarme.

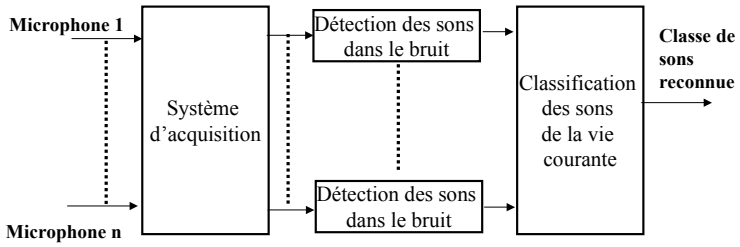


Fig. 2.1: Structure du système d'extraction d'informations sonores

Après avoir conçu la structure générale du système pour trouver les algorithmes les mieux adaptés pour la détection et la classification de sons, nous devons étudier les caractéristiques des sons. La nature de l'application, télésurveillance des personnes âgées et/ou malades, impose le type de sons à étudier. Le but principal du système d'analyse sonore est de compléter un système équipé avec des capteurs physiologiques (mesure du pouls, de l'oxymétrie, du poids, etc.) pour la détection des situations de détresse et l'identification de diverses pathologies. Un corpus de sons spécifique à l'application est nécessaire pour valider les divers algorithmes qui composent les parties du système. Chaque type de signal doit être présent dans le corpus avec plusieurs répétitions pour permettre sa modélisation statistique.

Au début de nos recherches, il n'existait pas dans le milieu scientifique de corpus de sons de la vie courante approprié à des applications de télésurveillance médicale. Nous avons dû réaliser un tel corpus en enregistrant des sons et en récupérant des sons de CD commerciaux.

Le corpus contient aussi bien des sons associés à une situation de détresse comme des chutes d'objets, des bris de verre, des cris mais aussi des sons courants comme des claquements de porte, des sons de vaisselle, des sons de l'écoulement d'eau, des sons de pas, des sons de serrure de porte. Le corpus doit inclure le bruit environnemental de l'appartement et d'autres bruits considérés comme bruits environnementaux : bruit de l'écoulement de l'eau, bruit de sèche-cheveux, bruit de rasoir électrique, etc.

Ainsi, la composition du corpus de sons de la vie courante (Figure 2.2) qui a été crée est la suivante :

- 15% de sons enregistrés dans le studio ou à la cafétéria du CLIPS ;
- 15% de sons provenant d'un CD d'effets pour films [Sciascia, 1992] ;
- 70% de sons récupérés et transformés du CD «Sound Scene Database in Real Acoustical Environments»(RWCP) du laboratoire ATR [Partnership, 2001].

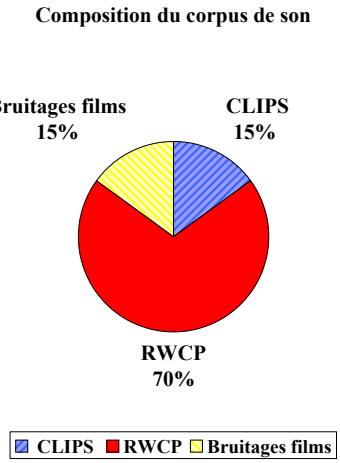


Fig. 2.2: Composition du corpus de sons de la vie courante

La méthodologie de réalisation du corpus s'est inspirée de celle utilisée dans les corpus de parole en ce qui concerne les fichiers d'étiquetages et de description. Dans la plupart des cas les corpus de parole sont enregistrés à une fréquence d'échantillonnage de 16 kHz. Le corpus a été enregistré à une fréquence d'échantillonnage de 44.1 kHz et sans compression même si notre système fonctionne à une fréquence d'échantillonnage de 16 kHz. Ce corpus pourra servir à d'autres applications. Pour le format des fichiers nous avons eu le choix entre le format «wav», qui s'est imposé sur les PC et un format de type «raw» (fichier binaire sans en-tête). Nous avons choisi le format «wav» parce qu'il permet la lecture par la majorité des logiciels et la conversion vers d'autres formats est facile. Le rapport signal sur bruit des enregistrements doit être au minimum de 70 dB pour pouvoir considérer les sons comme «purs». Le réglage de l'amplitude d'enregistrement est difficile dans le cas des sons impulsionnels qui ont une amplitude forte dans un intervalle réduit de temps.

Les enregistrements, réalisés dans le laboratoire CLIPS ont été faits sur une platine cassette numérique, en utilisant une fréquence d'échantillonnage de 48 kHz et un microphone Beyer Dynamics[1] de bande passante 20-20000 Hz. Le microphone utilisé présente une caractéristique spatiale de type cardioïde qui permet la capture du son désiré seulement. Un enregistrement du bruit environnemental de l'appartement d'étude a aussi été effectué. L'enregistrement de ce bruit que l'on va appeler «*bruit HIS*[2]» est fait avec le système d'acquisition et les microphones installés dans l'appartement. Ces microphones sont de type condensateur avec une caractéristique omnidirectionnelle.

Les signaux provenant du CD de laboratoire ATR ont une fréquence d'échantillonnage de 44.1 kHz, de même pour ceux du CD d'effets pour les films. Pour l'homogénéité du corpus les signaux enregistrés dans le studio CLIPS ont une fréquence d'échantillonnage de 44.1 kHz. Comme le système d'acquisition du son de l'appartement expérimental a une fréquence d'échantillonnage de 16 kHz, le corpus de sons contient aussi une version de chaque signal échantillonnée à 16 kHz obtenue par sous-échantillonnage des signaux de 44.1 kHz pour les amener à une fréquence d'échantillonnage de 16 kHz (le sous-échantillonnage étant d'une valeur non entière, est fait avec interpolation sur 5 échantillons).

Le corpus de sons de la vie courante est constitué de 3354 signaux pour une durée totale d'une heure et 35 minutes (≈650 Mo) et a été gravé sur le CD-Rom «Base de données. Sons de la vie courante» [Équipe GEOD Dan Istrate, 2001]. Parmi les sons du corpus nous avons : claquements de porte (différents types de porte), des sons de pas, des sons de rasoir électrique, des sons de sèche-cheveux, des sons de serrure de porte, des sons de vaisselle, des sons de bris de verre, des cris, des sons d'écoulement d'eau, des sons de sonneries, et aussi 3 minutes d'enregistrement du bruit HIS. En moyenne, le corpus contient 20 types de sons avec un minimum de 10 répétitions et un maximum de 300 répétitions pour chaque type.

Notre approche méthodologique, pour développer le système d'analyse sonore consiste à adapter les méthodes et les paramètres utilisés en reconnaissance de la parole/du locuteur aux particularités des sons de la vie courante. En effet, par exemple on peut constater sur le spectrogramme d'un signal de claquement de porte (figure 2.3) qu'il présente une large bande fréquentielle mais une courte durée (0.2s). Par opposition, sur le spectrogramme d'un signal de parole (exemple figure 2.4) la présence des fréquences privilégies (formants) [Boite et al., 2000] peut être constatée. Parmi les sons de la vie courante il y a aussi des sons qui présentent des fréquences fondamentales comme la sonnerie de téléphone (voir figure 2.5).

Pour tout le document, nous considérerons comme **sons utiles** (des sons impulsionnels et courts) : claquements de porte, bris de verre, sons de chutes d'objets, sons de pas, expressions parlées courtes, toux, éternuements, serrure de porte, sonneries et comme **bruit environnemental** (des bruits pseudo-stationnaires et de longue durée) : l'écoulement de l'eau, sèche-cheveux, rasoir électrique, ventilateur, etc.

Nous conviendrons d'utiliser les mots *son* pour **son utile** et *bruit* pour **bruit environnemental**.

[1] Modèle M260

[2] Habitat Intelligent Santé

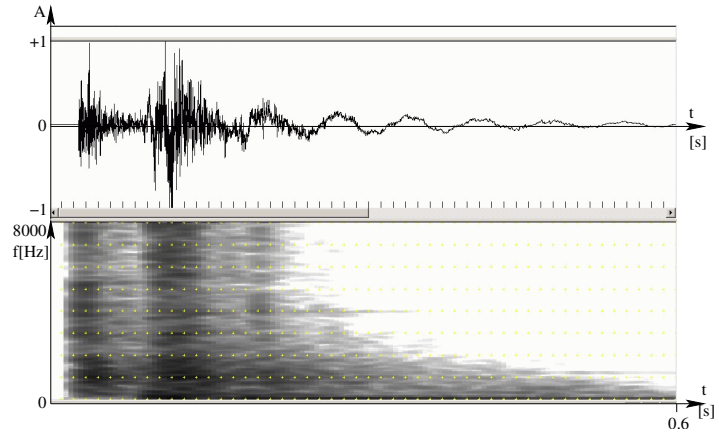


Fig. 2.3: Évolution temporelle et spectrogramme d'un claquement de porte

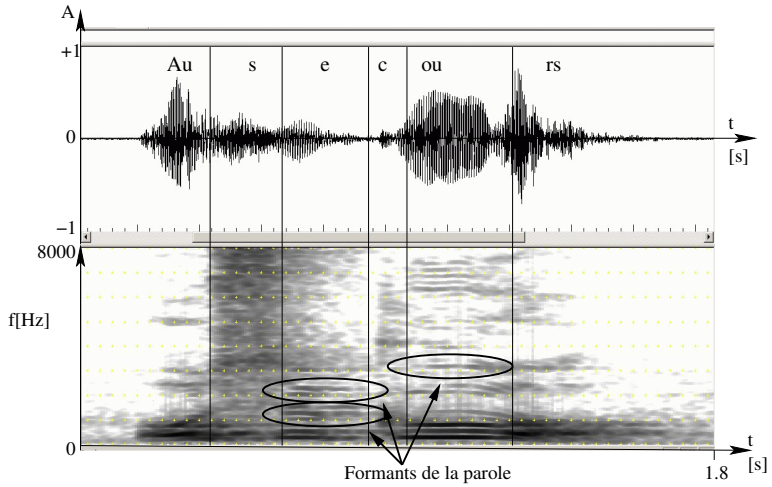


Fig. 2.4: Évolution temporelle et spectrogramme de l'expression «Au secours !»

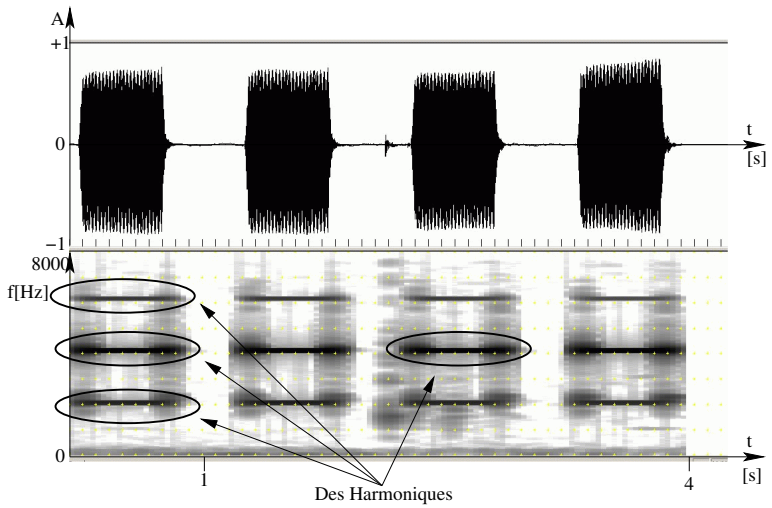


Fig. 2.5: Évolution temporelle et spectrogramme d'un signal de sonnerie de téléphone

Le CD présenté ci-dessus [Équipe GEOD Dan Istrate, 2001] est le corpus des sons de la vie courante qui constitue la base de départ pour trois autres corpus servant à l'évaluation et à la validation des différentes parties du système développées dans cette thèse (voir figure 2.6). Ces corpus sont obtenus par mélange entre des sons et des bruits ; ils sseront décrits en détail par la suite. Les acronymes de ces corpus sont expliqués dans le tableau 2.1.

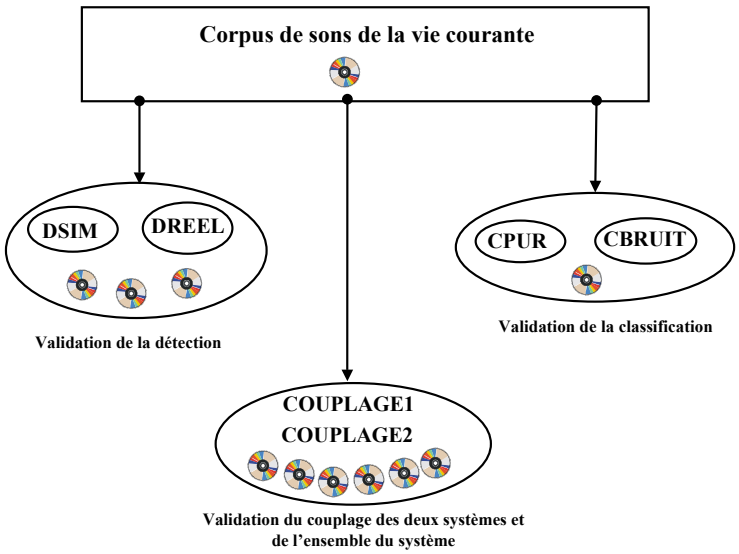


Fig. 2.6: Les corpus réalisé pour la validation des différentes parties du système

Acronyme	Type de corpus
DSIM	détection avec mélange simulé
DREEL	détection avec mélange réel
CPUR	classification avec sons purs
CBRUIT	classification avec sons bruités
COUPLAGE1	couplage, RSB variant entre 10 et 20 dB
COUPLAGE2	couplage, RSB variant entre 0 et 40 dB

Tab. 2.1: Acronymes des corpus de test

Le tableau 2.2 présente la composition du corpus de sons de la vie courante.

Son	Nbre de répétitions	Origine
Son agrafeuse	100	CD RWCP
Son d'applaudissement	390	CD RWCP
Son de déplacement de chaise	10	CLIPS
Sons humains (éternuements, bâillements, rires, ronflements, toux)	10	CD effets film
Sons d'ouverture récipient sous pression	100	CD RWCP
Son de chiffonnage ou déchirement de papier	200	CD RWCP
Son de pas	20	CD effets films(10) et CLIPS(10)
Son de perforeuse	100	CD RWCP
Claquements de différentes portes (porte d'entrée, d'armoire,de réfrigérateur)	548	CLIPS
Son de rasoir électrique	100	CD RWCP
Son de sèche-cheveux	100	CD RWCP
Son de serrure de porte	200	CD RWCP
Son de vaisselle	163	CLIPS(63) et RWCP(100)
Son de verre cassé	97	CD effets films
Son de chute de chaise, livre	420	CLIPS(20) et RWCP(400)
Cris	73	CD effets de film
Écoulement d'eau en lavabo, verre,douche	32	CLIPS(24) et CD effets film(8)
Sonneries de téléphone	520	CLIPS(30) et RWCP(480) et CD effets film(10)
Bruit de fond HIS	100	CLIPS

Tab. 2.2: Composition du corpus des sons de la vie courante

2.2 Corpus de sons pour la détection

2.2.1 Discussion sur le calcul du rapport signal sur bruit

Un paramètre très important pour mesurer la qualité du signal est le rapport signal sur bruit. Le rapport signal sur bruit est le rapport en dB entre la puissance moyenne du signal et celle du bruit ou, autrement dit, le rapport signal sur bruit est le rapport, en dB, entre la somme moyennée des carrés des échantillons du signal et la somme moyennée des échantillons du bruit. L'équation (2.1) donne la formule de calcul du rapport signal sur bruit.

$$RSB = 10 \cdot \log\left(\frac{P_{\text{signal}}}{P_{\text{bruit}}}\right) = 10 \cdot \log\left(\frac{\frac{1}{N}\sum_{i=0}^{N-1} s_i^2}{\frac{1}{M}\sum_{i=0}^{M-1} b_i^2}\right) \tag{2.1}$$

où :

- P_{signal} est la puissance moyenne du signal
- P_{bruit} est la puissance moyenne du bruit
- N le nombre d'échantillons du signal
- s_i les échantillons du signal
- M le nombre d'échantillons du bruit
- b_i les échantillons du bruit

Généralement, si le bruit est stationnaire, sa puissance moyenne est constante dans le temps. La puissance moyenne des signaux impulsionnels varie rapidement dans le temps. La mesure du rapport signal sur bruit (RSB) d'un son impulsionnel est donc difficile parce que l'énergie d'un tel son varie rapidement dans le temps. La meilleure approximation sera le calcul de la valeur instantanée du rapport signal sur bruit. Pour notre application, nous pouvons utiliser une approximation de la puissance moyenne du signal : soit sa valeur maximale (au moment du maximum), soit une moyenne sur toute la longueur du son. Dans ce document, le RSB est calculé sur toute la longueur du son impulsionnel conformément à l'équation (2.2)). La puissance moyenne du bruit est calculée sur le nombre N d'échantillons du signal et non sur le nombre d'échantillons du bruit comme dans l'équation (2.1).

$$RSB = 10 \cdot \log \left(\frac{\sum_{i=0}^{N-1} s_i^2}{\sum_{i=0}^{\mathbf{N}-1} b_i^2} \right) \tag{2.2}$$

En conclusion, pour des signaux impulsionnels le calcul exact du RSB n'est pas possible et seule une approximation de celui-ci peut être établie. L'équation (2.2) tient compte de la moyenne de l'énergie du signal en rapport avec celle du bruit sur la même durée.

2.2.2 Corpus pour la détection en conditions simulées - DSIM

Pour tester les algorithmes de détection d'événements sonores un corpus de test a été généré (DSIM), à partir du corpus de sons de la vie courante décrit précédemment. Le corpus contient pour chaque signal à détecter un fichier sonore d'une longueur de 25s. Cette longueur a été choisie en tenant compte de la longueur maximale des sons à détecter et du temps d'initialisation nécessaire pour certaines algorithmes $\approx$ 5s. Pour faciliter l'évaluation, l'instant d'apparition du signal utile est toujours le même et a été fixé à 10s à partir du début. Cette valeur est liée au temps d'initialisation des algorithmes et a été fixée plus grande que le temps maximal d'initialisation des algorithmes de détection pour avoir une marge (pouvoir tester des algorithmes qui nécessiteraient un temps d'initialisation plus long). Ce temps d'initialisation est négligeable tenant compte du fait que le système fonctionne sans arrêt 24h/24h. Les algorithmes de détection présentés dans le chapitre suivant se fondent sur des paramètres statistiques (comme la moyenne ou l'écart-type) nécessitant le remplissage d'un tampon de données. Cela explique la présence d'un temps d'initialisation de l'algorithme.

Nous avons utilisé comme fond sonore le bruit blanc, le bruit de l'écoulement de l'eau et le bruit environnemental de l'appartement d'étude appelé bruit HIS. Le bruit blanc a été choisi

pour faciliter la comparaison des résultats avec des algorithmes de l'état de l'art. Le bruit HIS est le bruit environnemental, donc il représente les conditions réelles et le bruit d'écoulement d'eau est un bruit statistiquement très présent dans la vie courante (lavage de la vaisselle, lavage du linge, douche, etc.).

Les différents sons à détecter, au total 11 types de sons, sont : cris, chutes d'une chaise, chutes d'un livre, bris de verre, claquements porte, sons de pas, toux, éternuements, serrures de porte, sonneries de téléphone et parole. Tous les fichiers sont des fichiers mono-canal, échantillonnés à 16KHz qui est la fréquence de travail du système d'acquisition sonore de l'appartement.

Nous avons décidé d'utiliser 4 valeurs pour le rapport signal sur bruit (RSB) : 0, 10, 20 et 40dB. La valeur maximale du rapport signal sur bruit est limitée par les qualité du système d'acquisition du son à une valeur proche de 40 dB. D'autre part la présence d'un rapport signal sur bruit plus petit que 0 dB n'a pas été envisagé ici, 0 dB représentant déjà une dégradation très importante pour un système de classification automatique.

Dans le cadre de la génération de notre base de test nous adaptons l'énergie du bruit pour obtenir le rapport signal sur bruit désiré [Dufaux, 2001]. Les différentes étapes de génération des fichiers de tests sont alors :

- Le calcul de l'énergie moyenne par échantillon du signal avec :

$$(E_{\text{signal utile}})_{dB} = 10 \cdot \log\left(\frac{1}{N}\sum_{i=0}^{N-1} s_i^2\right) \tag{2.3}$$

- La détermination du niveau de l'énergie moyenne du son nécessaire pour obtenir le rapport signal sur bruit (RSB) désiré sachant que :

$$E_{\text{bruit nécessaire}} = 10^{\frac{\left(E_{\text{signal utile}}\right)_{dB} - RSB}{10}} \tag{2.4}$$

- Le calcul de l'énergie moyenne par échantillon du bruit (sur le nombre d'échantillons du signal) suivant la formule :

$$E_{\text{bruit}} = \frac{1}{N}\sum_{i=0}^{N-1} b_i^2 \tag{2.5}$$

- Le calcul du coefficient de multiplication de chaque échantillon de bruit en vue d'obtenir le RSB désiré :

$$\text{Coeff} = \sqrt{\frac{E_{\text{bruit nécessaire}}}{E_{\text{bruit}}}} \tag{2.6}$$

Nous avons aussi introduit un coefficient dénoté «Coeff. anti-dépassement» dans l'équation (2.7), pour éviter le dépassement des limites du signal par la somme des échantillons du son avec ceux du bruit (coefficient qui tient compte des valeurs maximales du bruit et du son). Dans l'équation (2.7) $A_{\text{max bruit}}$ est l'amplitude maximale du bruit et $A_{\text{Max signal}}$ est l'amplitude maximale du signal. Les coefficients «Coeff» et «Coef. anti-dépassement» sont déterminés sur toute la durée du signal.

$$\text{Coeff. anti-dépassement} = \frac{0.95}{A_{\text{max bruit}} + A_{\text{Max signal}}} \tag{2.7}$$

La figure 2.7 présente l'addition (melange) d'un signal de claquement de porte avec un bruit environnemental de type HIS avec un RSB de 0dB. La définition du RSB, donnée dans la section précédente, explique le fait que l'amplitude maximale du son n'est pas égale à celle du bruit. En fait, c'est l'énergie moyenne du son qui est égale à celle du bruit pour la même période temporelle.

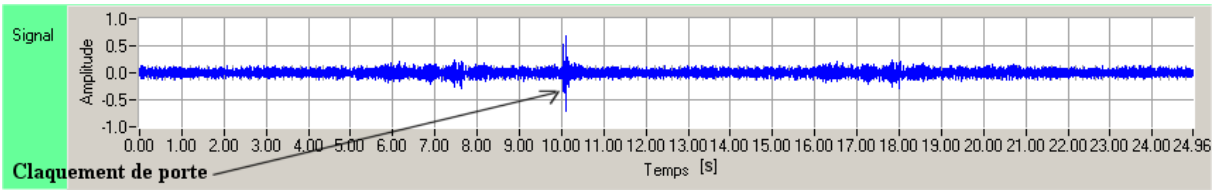


Fig. 2.7: Un exemple de son du corpus de détection (claquement de porte superposé au bruit HIS avec un RSB de 0dB

Nous avons utilisé pour chaque type de bruit et chaque type de son trois enregistrements différents. Nous disposons donc de 11 classes de signaux utiles avec trois enregistrements chacun, trois types de bruits de fond avec, là aussi, trois enregistrements chacun (figure 2.8) ; nous modifierons les enregistrements des bruits de fond pour obtenir les 4 valeurs de rapport signal sur bruit prévues avant de les additionner aux signaux utiles. Cela conduit à $11 * 3 * 3 * 3 * 4 = 1188$ fichiers. Nous avons aussi généré un nombre identique de fichiers comportant seulement le bruit (sans son), ce qui fait que notre base comporte au total 2376 fichiers (≈ 3 CD de 650 Mo chacun, un CD par type de bruit environnemental). Le tableau 2.3 conclut sur la composition de ce corpus. Les fichiers sans signal utile ont été générés pour tester la réponse du système seulement sur le bruit environnemental.

Nbre de fichiers	Durée	Nbre de signaux à détecter
1188 fichiers avec signal utile	8h 15min	1188
1188 fichiers sans signal utile	8h 15min	0
TOTAL : 2376 fichiers	16h 30min	1188

Tab. 2.3: Principales caractéristiques du corpus de sons pour la détection

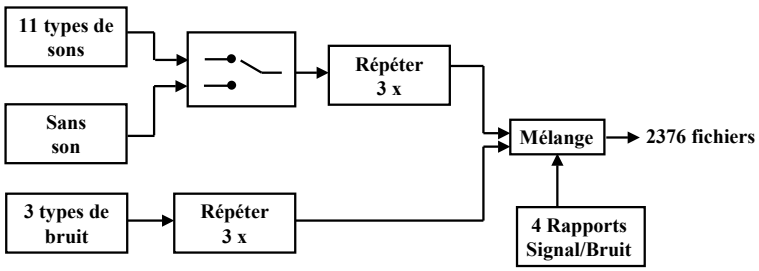


Fig. 2.8: Procédure d'obtention du corpus de sons pour la détection

Pour chaque fichier sonore du corpus, il y a un fichier de type SAM [Wells et al., 1992] qui marque par des étiquettes le début et la fin du signal à détecter. Un fichier de type SAM

est un fichier texte d'étiquetage des fichiers sonores. Ces fichiers sont utilisés pour calculer les performances des algorithmes de détection d'événements sonores.

2.2.3 Corpus pour la détection en conditions réelles - DREEL

La validation finale des meilleurs algorithmes de détection sera faite sur un petit corpus de sons (DREEL) enregistrés en conditions réelles. Il contient les mêmes 11 signaux du corpus principal pour la détection reproduits dans l'appartement d'étude avec une chaîne constituée d'un ordinateur attaquant un amplificateur professionnel (Yamaha[3]) équipé de haut-parleurs (Elipson[4]). Au total, nous avons 60 fichiers avec des rapports signal sur bruit qui varient entre 2 dB et 30 dB. Le RSB en moyenne est de 15 dB. La longueur des fichiers est aussi de 25 secondes. La différence par rapport au corpus précédent consiste dans le fait que le melange bruit HIS - son est réel et l'enregistrement est effectué par la chaîne d'acquisition du système expérimental, à l'intérieur de l'appartement (plus exectement dans le séjour et dans le couloir parce que l'echo est réduit).

Ce corpus servira à valider sur le corpus en conditions réelles les résultats obtenus avec des mélanges produits par simulation.

Sur ce corpus nous avons effectué une mesure de la repartition des valeurs du RSB dans l'appartement d'étude qui a donnée l'histogramme de la figure 2.9.

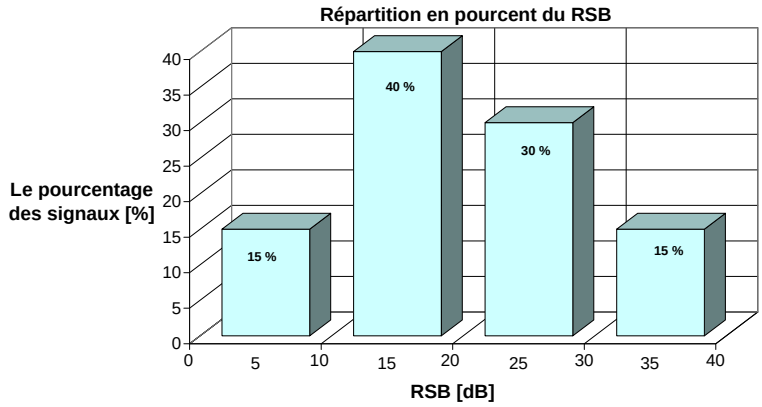


Fig. 2.9: Répartition en pourcentage des rapports signal sur bruit des signaux enregistrés dans l'appartement HIS

2.3 Corpus de sons pour la classification

2.3.1 La composition du corpus pur - CPUR

Les sons de la vie courante sont très variés et avec une importance différente pour une application de télésurveillance médicale. Pour notre application sont considérés comme intéressantes les sons de la vie courante suivants :

- sons indiquant une situations de détresse : cris, chute, bris de verre, sons de vaisselle.

[3] Amplificateur Yamaha AX892 270W
[4] Enceinte Elipson MINIB 60W

- sons qui peuvent identifier une pathologie : toux, éternuement, sons de pas, chasse d'eau. Par exemple, les sons des pas et de l'eau pourraient être intéressants pour détecter des troubles urinaires nocturnes.
- sons aidant une localisation de la personne : sons de pas, claquement de porte, etc.
- sons nécessaires au système pour l'élimination des fausses alarmes : sonnerie de téléphone, sonnerie de porte, etc. La détection d'une sonnerie de téléphone indiquera au système une possible conversation téléphonique et donc, réduire la confiance des éventuelles expressions de détresses reconnus.

Le corpus de sons pour la classification (CPUR) a été limité dans une première étape à 7 classes. Des sons comme la chute de personnes ou d'objets, les toux sont difficiles à reproduire et n'ont pas encore été enregistrés. Pour l'application médicale, les sons de longue durée et stationnaires n'apportent pas d'information et sont considérés comme bruit environnementaux. Le tableau 2.4 montre les classes de sons qui sont divisées en deux catégories :

- Classes de sons normaux ($\overline{A}$) comme : claquements de porte, sonneries de téléphone, sons de pas, sons humains, serrures de porte.
- Classes de sons qui génèrent une alarme (A) : bris de verre, cris, chute, vaisselle

La classe de son	Nbre de fichiers	Nbre de trames	Durée totale	Alarme
C1 - Claquement de porte	523	47398	474 s	NON
C2 - Bris de verre	88	9338	93 s	**OUI**
C3 - Sonneries de téléphone	517	59188	592 s	NON
C4 - Sons de pas	13	36480	365 s	NON
C5 - Cris	73	17509	175 s	**OUI**
C7 - Vaisselle	163	7943	79 s	**OUI**
C9 - Serrures de porte	200	6050	61 s	NON

Tab. 2.4: Les 7 classes de sons et le nombre de fichiers et de trames par classe

Dans le même tableau (tableau 2.4) nous pouvons observer le nombre des signaux par classe de sons, le nombre des trames et la durée totale des signaux de chaque classe. Le nombre de trames de chaque classe est très différent d'une classe à une autre. Pour une modélisation statistique des sons, il y a un nombre minimal des trames nécessaires à une bonne représentation.

2.3.2 Le corpus bruité - CBRUIT

Pour tester et valider le système de classification des sons nous avons construit une base de tests bruitée. Dans cette base, pour chaque classe de sons, il y a quatre mélanges entre le son et le bruit HIS correspondant aux quatre valeurs de RSB : 0, 10, 20 et 40 dB. Les mêmes valeurs de RSB utilisées dans le corpus de sons pour la détection ont été conservées.

L'amplitude du bruit HIS est calculée et adaptée en fonction de la valeur du RSB désirée comme pour le corpus de sons de détection (voir section 2.2).

Ce corpus a été créé seulement en utilisant le bruit HIS parce que c'est le bruit réel de l'appartement, le bruit qui sera présent tout le temps dans les signaux capturés. Tenant compte

du grand nombre de tests dans une première étape, nous avons éliminé le bruit blanc (qui n'est pas réaliste) et le bruit de l'écoulement de l'eau.

Ce corpus permet l'étude de l'influence du bruit sur le système de classification dans les conditions d'une segmentation parfaite des sons. Ce corpus sert aussi, pour tester et valider la deuxième solution proposée : le débruitage du signal entre la phase de détection et celle de classification.

2.4 Corpus pour le couplage entre la détection et la classification - COUPLAGE1, COUPLAGE2

Par la suite, nous avons été amenés à créer un dernier corpus de test qui contient tous les signaux des classes de sons utilisés pour la classification, superposés au bruit HIS. Les signaux d'une classe de sons s'enchaînent sur un même fichier sonore. Ce corpus sert à tester et valider le système complet composé par la détection et la classification des sons de la vie courante. Chaque fichier sonore est constitué d'une succession de signaux de la classe correspondante entre lesquels sont intercalés des intervalles de silence, de durée aléatoire (Figure 2.10). Tous les sons sont mélangés avec le bruit HIS à un rapport signal sur bruit qui varie aléatoirement, soit d'après une loi uniforme dans l'intervalle [10 - 20] dB, soit d'après une loi pseudo-gaussienne entre [0-40] dB. La loi de variation du deuxième cas a été définie, après avoir effectué une statistique des rapports signal sur bruit des signaux acquis dans l'appartement d'étude (figure 2.9).

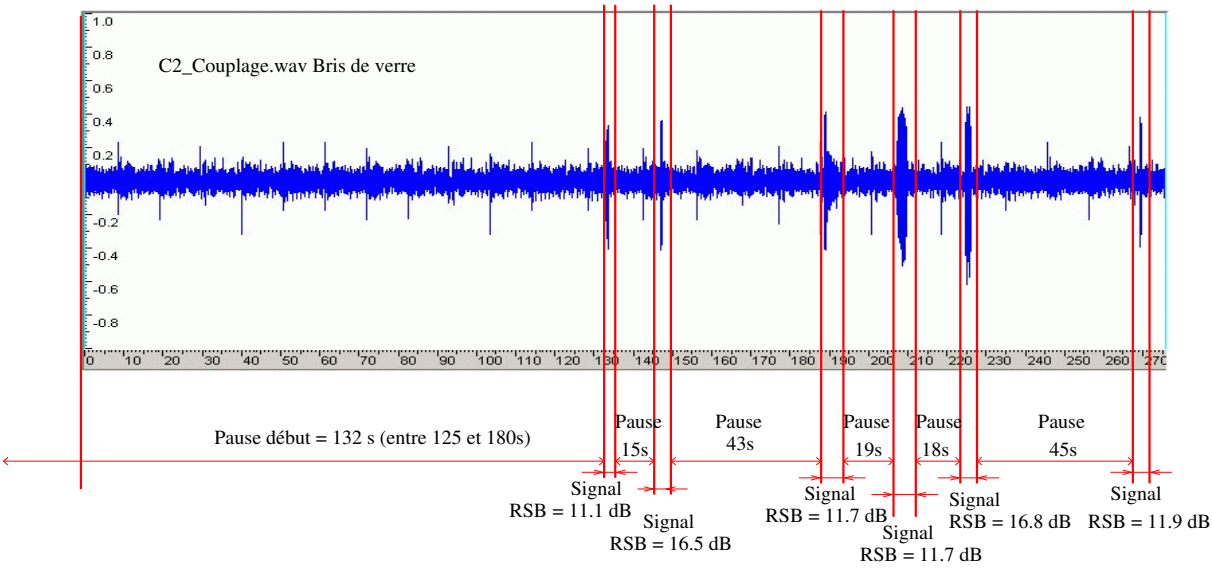


Fig. 2.10: Un exemple du début du fichier correspondant à la classe de sons des cris

Les paramètres caractéristiques à chaque fichier sont :

- la durée du silence du début de fichier varie en mode aléatoire entre 2 minutes 5 secondes et 3 minutes
- la durée du silence entre deux signaux varie en mode aléatoire entre 5 et 60 secondes
- le rapport signal sur bruit du fichier varie, pour chaque signal, en mode aléatoire :

- ***1er cas*** - uniformément entre 10 et 20 dB
- ***2e cas*** - entre 0 et 40 dB en respectant la répartition statistique présentée dans la figure 2.9

L'amplitude du bruit HIS est constante et égale à la valeur du système d'acquisition expérimental. L'amplitude du son est adaptée en fonction de la valeur du rapport signal sur bruit désirée, valeur générée en mode aléatoire pour chaque signal. Après le tirage aléatoire de la valeur du rapport signal sur bruit pour chaque son à insérer, nous calculons le coefficient de correction avec la formule 2.8. Ce coefficient multiplie chaque échantillon du son avant la réalisation du mélange avec le bruit. Nous avons, aussi, introduit un coefficient pour éviter le dépassement des limites du signal par la somme des échantillons du signal et ceux du bruit (coefficient qui tient compte des valeurs maximales du bruit et du son). Dans l'équation (2.9) $A_{\text{max bruit}}$ est l'amplitude maximale du bruit et $A_{\text{Max signal}}$ est l'amplitude maximale du signal.

$$\text{Coeff. correction} = \sqrt{\frac{10^{\frac{E_{Bruit}+RSB}{10}}}{E_{\text{signal utile}}}} \tag{2.8}$$

$$\text{Coeff. anti-dépassement} = \frac{0.95}{\text{Coeff. correction} * A_{\text{max bruit}} + A_{\text{Max signal}}} \tag{2.9}$$

La répartition statistique du tirage aléatoire est uniforme quand le rapport signal sur bruit est inclus dans l'intervalle [10-20] dB. Par contre, quand le rapport se situe dans l'intervalle [0 - 40] dB, le tirage tient compte de la répartition statistique présentée dans la figure 2.9 (de la page 44). Cette variation discrète de la densité de probabilité du RSB represente l'histogramme de répartition des RSB pour les signaux enregistrés dans l'appartement HIS.

Pour chaque fichier sonore, il y a un fichier de type SAM [Standard SAM, 1992] qui marque par des étiquettes le début et la fin de chaque signal à détecter, l'identificateur du signal et le rapport signal sur bruit de chaque signal. Ces fichiers sont utilisés pour calculer les performances des algorithmes de détection d'événements sonores.

La taille du corpus avec le RSB qui varie uniformément dans l'intervalle de 10-20 dB est de 13 heures 8 minutes et celle du corpus avec le RSB variant entre 0 et 40 dB est de 14 heures 14 minutes (voir tableau 2.5). La différence de taille provient des intervalles de silence qui ont une durée aléatoire. Le nombre total des signaux à détecter et reconnaître est de 1577. Le nom des fichiers correspond aux noms des classes sonores du corpus de classification.

Nom du fichier	COUPLAGE1 $RSB \in [10 - 20]\ dB$		COUPLAGE2 $RSB \in [0 - 40]\ dB$	
	Durée	Nbre de signaux à détecter	Durée	Nbre de signaux à détecter
C1	16711 s	523	16707 s	523
C2	2749 s	88	2836 s	88
C3	14639 s	517	14639 s	517
C4	648 s	13	534 s	13
C5	2353 s	73	2459 s	73
C7	4944 s	163	5351 s	163
C9	5233 s	200	6458 s	200
Total	**13h 8min**	**1577**	**14h 13min**	**1577**

Tab. 2.5: Nombre des signaux et durées des fichiers du corpus de sons pour le couplage

Ce corpus sert, principalement, à étudier l'influence d'une segmentation imparfaite sur le système de classification de son. Par «segmentation imparfaite» nous entendons une identification approximative du début du son, suivie dans un premier temps de l'extraction d'un segment de durée fixe du signal. Dans le chapitre 5, une méthode d'identification de la fin du signal pour améliorer les performances de classification sera néanmoins proposée. Ce corpus sert aussi à l'évaluation de l'ensemble du système.

2.5 Conclusions

Ce chapitre a présenté la creation d'un corpus des sons de la vie courante approprié à une application de télésurveillance médicale. Le corpus de sons contient ≈3300 signaux échantillonés à 44.1 kHz et 16 kHz (la fréquence d'échantillonage du système d'acquisition implementé dans l'appartement d'étude). Ce corpus de sons a servi à la réalisation des trois autres corpus nécessaires à l'évaluation et à la validation des différents étapes du système proposé : détection, classification et couplage entre les deux.

Bibliographie

[Boite et al., 2000] Boite, R., Bourlard, H., Dutoit, T., Hang, J., and Leich, H. (2000). *Traitement de la parole*. ISBN 2-88074-388-5. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Dufaux, 2001] Dufaux, A. (2001). *Detection and recognition of Impulsive Sounds Signals*. PhD thesis, Faculté des sciences de l'Université de Neuchâtel, Suisse.

[Partnership, 2001] Partnership, R. W. C. (1998-2001). CD - Sound scene database in real acoustical environments. http ://tosa.mri.co.jp/sounddb/indexe.htm.

[Équipe GEOD Dan Istrate, 2001] Équipe GEOD Dan Istrate, C.-I. (2001). Base de données. Sons de la vie courante.

[Sciascia, 1992] Sciascia, S. (1992). CD - bruitages, vol.3.

[Standard SAM, 1992] Standard SAM (1992). http ://www.icp.grenet.fr/relator/standsam.html.

[Wells et al., 1992] Wells, D., Barry, J., Grice, W., Fourcin, M., and Gibbon, A. (1992). SAM ESPRIT PROJECT 2589 - multilingual speech input/output assessment, methodology and standardisation. Final report. Technical Report SAM-UCL-G004, University CollegeLondon.

Chapitre 3

Détection des sons dans le bruit

3.1 Objectifs

Ce chapitre est consacré à l'étude de la détection d'événements sonores ; la détection constitue la première phase du système d'extraction d'informations par analyse sonore. Le rôle du système de détection est de déterminer l'instant d'apparition d'un événement sonore, en vue de l'extraire du bruit de fond pour un traitement ultérieur de classification. Nous considérons comme événement sonore les sons de la vie courante ou la parole et dans la plupart des cas le bruit de fond est un bruit stationnaire. La position de cet étage dans le cadre de l'application sonore est présentée en gras dans la Figure 3.1. La détection des événements sonores est effectué indépendamment sur chaque canal sonore.

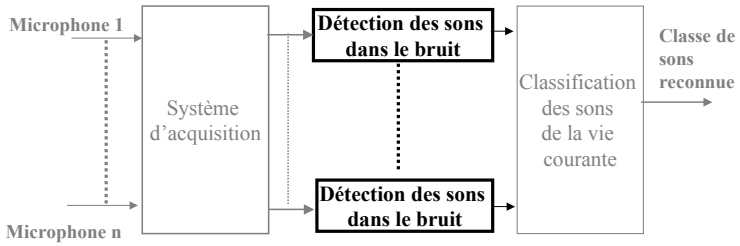


Fig. 3.1: Structure du système d'extraction d'informations sonores

Détecter consiste à identifier l'instant d'apparition d'un signal recherché dans un environnement bruité. En théorie de la détection binaire, le problème de détection revient à décider de la présence ou de l'absence du signal dans le bruit. Chacune de ces situations peut se formaliser comme une hypothèse H_0 ou H_1 (voir équation 3.1). L'algorithme de détection teste l'hypothèse H_0 contre l'hypothèse H_1 pour décider laquelle est vraie. Nous définissons les hypothèses de la façon suivante :

$$\begin{cases} H_0 & o(t) = b(t) \\ H_1 & o(t) = s(t) + b(t) \end{cases} \tag{3.1}$$

en notant $o(t)$ l'observation du signal analysé, $b(t)$ le bruit additif et $s(t)$ le signal à détecter.

On associe généralement à un problème de détection une fonction appelée *test de détection* qui dépend de l'observation $o(t)$. Cette fonction test $d(o)$ prend des valeurs binaires indiquant la décision prise comme défini dans l'équation 3.2).

$$\begin{cases} d(o) = 0 \Leftrightarrow \text{ si } o \in \mathcal{S}_0 \quad \text{on décide } H_0 \\ d(o) = 1 \Leftrightarrow \text{ si } o \in \mathcal{S}_1 \quad \text{on décide } H_1 \end{cases} \tag{3.2}$$

où $\mathcal{S}_0$ et $\mathcal{S}_1$ sont les deux partitions de l'espace complet d'observation $\mathcal{S}$ ($\mathcal{S} = \mathcal{S}_0 \bigcup \mathcal{S}_1$; $\mathcal{S}_0 \bigcap \mathcal{S}_1 = \emptyset$).

Le test de détection est toujours constitué de la comparaison d'une fonction $p(o)$ à un seuil λ, o désignant l'observation. Les deux partitions sont définies de la façon suivante :

$$\begin{cases} \mathcal{S}_0 = \{o \in \mathcal{S} | p(o) < \lambda\} \\ \mathcal{S}_1 = \{o \in \mathcal{S} | p(o) \geq \lambda\} \end{cases} \tag{3.3}$$

A chaque algorithme de détection sont associés une fonction $p(o)$ et un seuil λ qui correspondent à une erreur de détection minimale. Pour une détection binaire, les 4 cas possibles sont présentés dans le tableau 3.1.

Réponse du système	H_0 vraie	H_1 vraie
H_0	Non détection correcte	Détection manquée
H_1	Fausse alarme	Détection correcte

Tab. 3.1: Les situations possibles pour la détection

Les performances de la détection influencent les performances de l'ensemble du système. Par exemple un grand nombre de fausses alarmes sature le système de classification ; d'autre part, ne pas détecter un son peut avoir une conséquence grave. Une autre contrainte est le fonctionnement en temps réel. Le système doit fonctionner en temps réel sur plusieurs canaux (5 dans notre cas) en même temps, alors il doit être suffisamment rapide.

Généralement un système de détection doit calculer les paramètres du signal sur des échantillons passés, et avec ces valeurs et celles de la fenêtre d'analyse courante, obtenir un signal qui permettrait par un simple seuillage la détection d'un événement sonore (conformément au synoptique général présenté dans la figure 3.2). Le seuil appliqué peut être fixe ou adaptatif. Les signaux que nous voulons détecter ont une variation rapide d'énergie, une courte durée temporelle (sauf pour la parole) et une large bande de fréquences (plutôt hautes fréquences). Une approche simpliste serait donc, de suivre l'évolution de l'énergie et de réaliser un seuillage, mais cette méthode donne des bons résultats seulement pour des RSB élevés. En conclusion, nous devrons prendre en compte des variations plus complexes comme l'évolution de l'énergie ou la variation de la bande de fréquences du signal comme indicateurs de détection. En ce qui concerne le seuillage, une adaptation du seuil en fonction de l'évolution du signal semble indispensable pour rendre l'algorithme indépendant des gains du système d'acquisition.

La Figure 3.3 donne un exemple de son à détecter. Dans la partie supérieure de la figure est représenté le son d'une sonnerie de téléphone apparaissant entre la $10^{\text{ème}}$ et $11^{\text{ème}}$ seconde du signal, ce dernier étant noyé dans le bruit de l'écoulement d'eau. L'amplitude du signal est

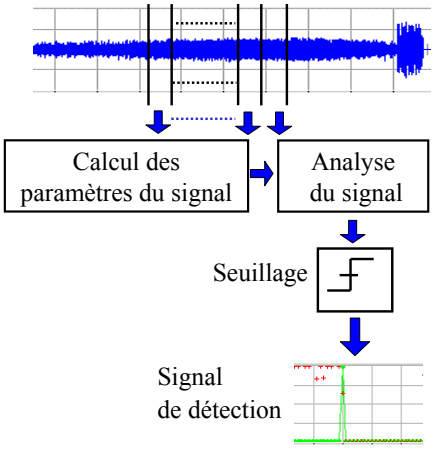


Fig. 3.2: Structure générale du système de détection

la même que celle du bruit de fond, avec un RSB de 0 dB. Il est donc difficile de le détecter avec un simple seuillage. Dans la partie inférieure de la *figure 3.3* est présentée l'énergie du signal et nous constatons que celle du signal à détecter ne dépasse pas celle du bruit de fond : on ne peut pas détecter non plus le signal par un simple seuillage sur l'énergie du signal [Dufaux et al., 2000].

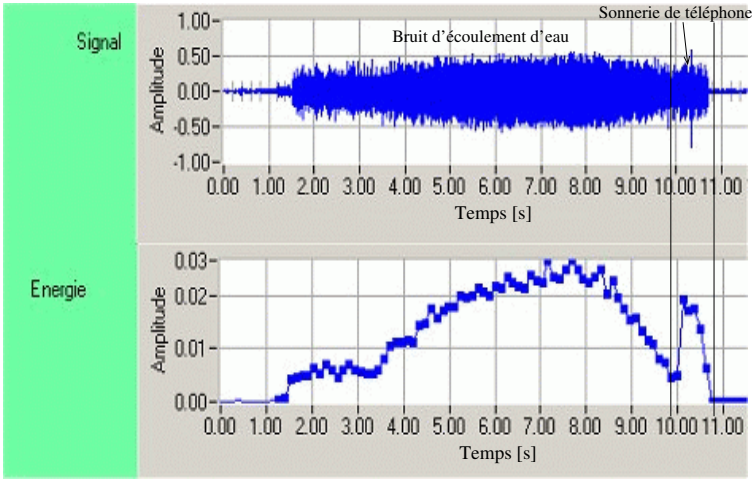


Fig. 3.3: Exemple d'un signal «noyé» dans un bruit d'écoulement de l'eau

Certaines techniques de détection des signaux impulsionnels utilisent l'énergie du signal et essaient, par traitement du signal (filtrage médian, calcul des paramètres statistiques, etc.), de faire ressortir les maxima locaux [Sohn et al., 1999]. Les résultats dépendent beaucoup des propriétés des bruits de fond. Parmi les méthodes de détection existantes il y a, aussi, celles de détection de la voix (VAD-Voice Activity Detection) mais celles-ci utilisent des propriétés spécifiques à la parole, comme par exemple, la présence de la fréquence fondamentale [Gökhun and Özer, 2000].

Une qualité principale du système de détection est sa *sensibilité* qui représente la capacité du système à détecter des événements sonores qui ont une amplitude proche ou égale à celle du bruit de fond. Une grande sensibilité permet la détection de tous les événements sonores

mais aussi des événements inexistants (les fausses alarmes). Une sensibilité réduite élimine les fausses alarmes mais rend possible la non détection des vrais événements sonores (les détections manquées). Le choix de la sensibilité du système est le résultat du compromis entre le nombre de fausses alarmes admises et le nombre de détections manquées admises.

Les détections manquées et les fausses alarmes ne sont pas les seules erreurs possibles. Il faut tenir compte aussi du manque de précision de la localisation du début d'événement. Pour notre système, nous avons choisi de considérer qu'une détection réalisée 0.5s avant le véritable début de l'événement et qu'une détection produite après la fin de l'événement sont des fausses alarmes. Cette valeur de 0.5s a été choisi tenant compte de la sensibilité du système de classification qui suit et de la structure des algorithmes de détection (voir le paragraphe 3.2.1.4).

Une autre propriété importante du système est la capacité d'adaptation au bruit de fond qui varie dans le temps (voir le cas de l'écoulement de l'eau de la Figure 3.3). Par exemple, si un sèche-cheveux est mis en route par le patient à surveiller alors notre système peut détecter cet événement, mais si pendant le fonctionnement du sèche-cheveux une chute se produit, le système doit aussi être capable de détecter le son de la chute.

Dans la section suivante, après une présentation des méthodes existantes, trois algorithmes de l'état de l'art sont détaillés avec leurs performances obtenues sur notre corpus de sons pour la détection (DSIM). Il s'agit d'un premier algorithme simple basé sur la variance de l'énergie normalisée du signal, et deux autres fondés sur le filtre médian conditionné. Trois nouveaux algorithmes avec leurs principes et performances sont alors proposés et présentés dans la section qui suit. Le chapitre 3 se termine par une comparaison de tous les algorithmes, dans le cas particulier des espaces perceptifs.

3.2 État de l'art de la détection des signaux impulsionnels

Le vaste domaine de la détection des signaux mono-dimensionnels inclut :

- La détection des signaux radar
- La détection des signaux sonar
- La détection des signaux numériques dans les transmissions radio par modulation de fréquence ou de phase
- La détection de la parole
- La détection des signaux environnementaux

Les signaux radar se situent dans la bande de hautes fréquences (3MHz - 100 GHz [Colin, 2002]) et sont représentés par la transmission et la détection en réception d'une impulsion. Les techniques utilisées pour le traitement des signaux radar sont spécifiques à leurs bandes de fréquence, ce qui empêche l'utilisation de ces techniques pour les événement sonores.

La bande de fréquence des signaux sonar [Bouvet, 1992] est [2,30] kHz et la technique généralement utilisée consiste à mesurer la phase des signaux sonar réceptionnés par l'antenne. En effet, dans le cas du sonar actif, on procède à la mesure du temps de retard entre un signal émis par le système de sonar et l'écho correspondant obtenu. La détection se fait ainsi souvent sur la base d'une connaissance *a priori* du signal, ce qui n'est pas le cas dans la problématique traitée dans ce manuscrit.

La détection de signaux numériques dans le cadre des transmissions radio par modulation de fréquence ou de phase se résume à l'identification sur un signal binaire du niveau «zéro» logique ou «un» logique [Kirsteins et al., 1997]. En opposition, la détection des sons de la vie courante qui nous intéressent ici se fait sur des signaux audibles par l'oreille humaine (20 - 20000 Hz). Les méthodes de détection de la parole qui n'utilisent pas les caractéristiques fréquentielles de celle-ci pourraient être appliquées à la détection des signaux environnementaux [Gökhun and Özer, 2000]. La différence fréquentielle entre le signal de parole et ceux des sons de la vie courante est le résultat de la production différente de ces sons (pour la parole nous avons plusieurs modèles de production).

Historiquement, les premiers systèmes de détection de la parole dans le bruit (Voice activity detection - VAD) utilisaient une méthode de seuillage par énergie qui donne des bonnes performances pour des RSB élevés [Rabiner and Sambru, 1975]. Mais cette approche ne fonctionne pas pour des petits RSB parce que l'énergie du bruit environnemental devient égale à celle du son.

Parmi les méthodes classiques de détection par seuillage de l'énergie, l'adaptation du seuil à chaque analyse en utilisant la probabilité de distribution de l'amplitude du signal de parole est proposé par Özer dans [Ozer and Tanyer, 1998]. Le principe de cet algorithme, appelé GAET (Geometrically Adaptative Energy Threshold) est utilisable dans le cas de la détection des sons de la vie courante.

L'utilisation de la mesure de la périodicité de la parole est une méthode possible de VAD [Irwin, 1980], appelé LSPE (Least-Square Periodicity Estimator). Cette méthode présente des bonnes performances pour des RSB élevés mais n'est pas utilisable à des RSB réduits. Parmi les sons de la vie courante nous trouvons des sons qui présentent des périodicités mais la plus grande partie d'entre eux n'en ont pas.

La probabilité de distribution de l'amplitude des signaux de parole et du bruit est utilisée pour déterminer le niveau du bruit. L'estimation du niveau de bruit utilisé conjointement avec celui de l'énergie sert à détecter la présence de la parole. Tanyer propose dans [Tanyer and Ozer, 2000] la fusion entre cette méthode et la méthode LSPE. Les deux techniques utilisent les différences en termes de probabilité de distribution de l'amplitude et présence de la périodicité entre la parole et le bruit.

Une autre technique de détection de la parole est celle qui se fonde sur le nombre de passages par zéro pour différencier la parole du bruit. L'hypothèse de cette méthode est que le nombre de passages par zéro pour le bruit est considérablement plus grand que pour la parole [Junqua et al., 1991].

Un autre type de méthode utilise les statistiques d'ordres supérieures sur le résidu de la prédiction linéaire pour différencier la parole du bruit [Nemer et al., 2001]. Cette méthode se base sur le modèle de production de la parole (le son des cordes vocales est filtré par le conduit vocal) et n'est pas valable pour les sons de la vie courante. En conséquence cette méthode semble difficile à être appliquée pour la détection des divers sons de la vie courante noyés dans un bruit stationnaire.

L'utilisation des modèles statistiques pour la détection est une autre méthode possible [Sohn et al., 1999]. Par exemple, l'utilisation d'une règle de décision bayesienne [Zhang et al., 2002] donne de bons résultats de segmentation parole/bruit. Une méthode plus complexe est celle proposée par T.Yamada et N.Watanabe dans leur article ([Yamada and Watanabe, 2001]) qui utilise des modèles HMM du bruit environnemental et de

la parole pour détecter la présence de la parole.

La référence la plus proche de notre problème est la thèse de Alain Dufaux qui a étudié la détection et la classification de sons impulsionnels pour un système anti-effraction [Dufaux, 2001].

Tenant compte des contraintes et des techniques existantes, nous nous sommes focalisés sur les algorithmes fondés sur le traitement du signal comme l'analyse de l'énergie du signal ou de son spectre fréquentiel [Max, 1989], [Gargour and Samir, 2001]. Les algorithmes de détection par méthodes statistiques sont trop complexes et ne répondent pas aux contraintes de temps réel [Seck, 2001]. L'étude a commencée par l'adaptation et la validation de trois méthodes proposées dans [Dufaux, 2001].

3.2.1 Algorithmes étudiés issus de l'état de l'art

Pour avoir une première idée des performances, trois algorithmes de détection des sons impulsionnels, issus de l'état de l'art, ont été étudiés, adaptés et validés sur le corpus de sons (DSIM) réalisé spécialement pour la détection [Duvaut, 1991]. Le principe de base du premier est la détection d'une diminution de la variance de l'énergie normalisée en présence d'un son impulsionnel. Les deux autres sont fondés sur le filtrage médian (un filtrage non-linéaire) ou médian conditionné de l'énergie du signal analysé.

3.2.1.1 Algorithme fondé sur la variance

Cet algorithme très simple dans son principe, détecte l'apparition d'un signal impulsionnel en utilisant comme fonction $p(o)$ (voir introduction) la variance de l'énergie du signal. L'algorithme calcule l'énergie et réalise une normalisation de celle-ci sur une fenêtre de plusieurs trames, conformément à l'organigramme de la Figure 3.4. L'énergie du signal est calculée sur une tram de N échantillons. C'est cette normalisation de l'énergie (une transformation non-linéaire) qui permet de réaliser la détection de l'apparition d'un signal impulsionnel par simple seuillage de la courbe de variance du signal.

L'apparition d'un son impulsionnel produit une variation de courte durée de l'énergie du signal. Cette variation de courte durée change très peu la valeur de la moyenne de l'énergie à cause du nombre important des points utilisés pour le calcul de la moyenne et il en résulte alors une diminution de la variance de l'énergie du signal. Ces propriétés sont valables seulement si l'analyse statistique se fait sur un nombre suffisant d'échantillons du signal.

Les étapes de calcul de cet algorithme sont :

- Calcul de l'énergie moyenne par échantillon du signal à partir de la formule :

$$E_t = \frac{1}{N} \sum_{i=0}^{N-1} x_i^2 \tag{3.4}$$

  où x_i désigne un échantillon du signal et N le nombre d'échantillons par fenêtre d'analyse ;

- Normalisation des valeurs de l'énergie dans une fenêtre de longueur L avec :

$$E_{\text{normalisé}}(i) = \frac{E_i - \min\limits_j E_j}{\max\limits_j (E_j - \min\limits_j E_j)} \qquad \text{où} \qquad j = 0 \ldots L-1 \tag{3.5}$$

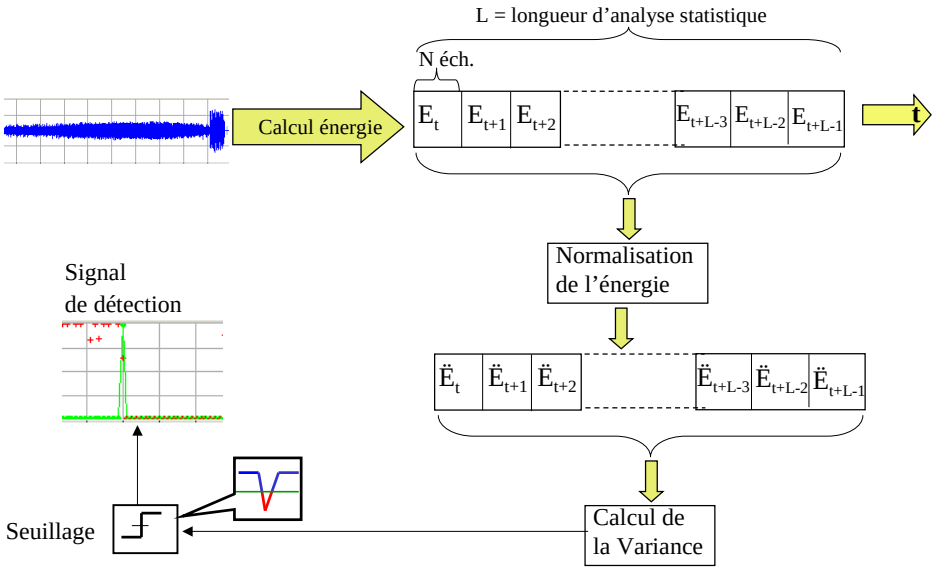


Fig. 3.4: Organigramme de l'algorithme fondé sur la variance de l'énergie

- Calcul non-biaisé de la variance de cette énergie normalisée :

$$p(o) = \sigma^2 = \frac{1}{L-1} \cdot \sum_{i=0}^{L-1} (E_{\text{normalisé}}(i) - \mu)^2 \qquad \text{où} \qquad \mu = \frac{1}{L} \cdot \sum_{i=0}^{L-1} E_{\text{normalisé}}(i) \quad (3.6)$$

- Seuillage de la variance obtenue : si la valeur de la variance descend au-dessous d'un seuil fixé (λ), on considère qu'un signal a été détecté.

$$d(\sigma^2) = \begin{cases} 1 \text{ si} & \sigma^2 \leq \lambda \\ 0 \text{ si} & \sigma^2 > \lambda \end{cases} \quad (3.7)$$

Le choix du nombre d'échantillons N et du nombre L de fenêtres d'énergie se fait en tenant compte des points suivants :

1. Le choix du nombre des échantillons N dans la fenêtre de calcul de l'énergie dépend des propriétés du signal et de la résolution temporelle nécessaire. Si N est trop petit, nous aurons beaucoup de détails dans la séquence de l'énergie, ce qui peut rendre difficile la détection. Par contre, une valeur trop grande peut conduire à des problèmes dans la détection des sons de courte durée (le plus court son a 250 ms). La valeur de N donne la précision théorique de la détection qui ne doit pas dépasser 0.5 s. En conclusion, nous avons choisi une valeur pour N de 2048 échantillons (une puissance de 2 pour permettre des calculs de TFR) qui donne une résolution temporelle[1] de 128 ms.
2. Le nombre L de fenêtres d'énergie sur lesquelles est effectuée la normalisation doit être assez grand pour que la variance calculée après la normalisation des valeurs de l'énergie soit représentative. La variance de l'estimateur σ^2 est donnée par l'équation 3.8.

$$Var(\sigma^2) = \frac{L-1}{L^3}[(L-1)\mathcal{E}[(E - \mathcal{E}(E))^4] - (L-3)\sigma^2] \quad (3.8)$$

[1] La fréquence d'échantillonnage pour tous les signaux est de 16 kHz

où $\mathcal{E}$ est l'espérance mathématique. L'algorithme a été évalué sur la base de tests pour différentes valeurs de L en obtenant la courbe présentée dans la figure 3.5. Nous avons choisi $L = 46$, ce qui correspond à $\approx 5.88s$ de temps d'initialisation de l'algorithme qui est négligeable pour une application de type espace perceptif. Une telle application, une fois démarrée, est censée fonctionner sans arrêt 24h/24h.

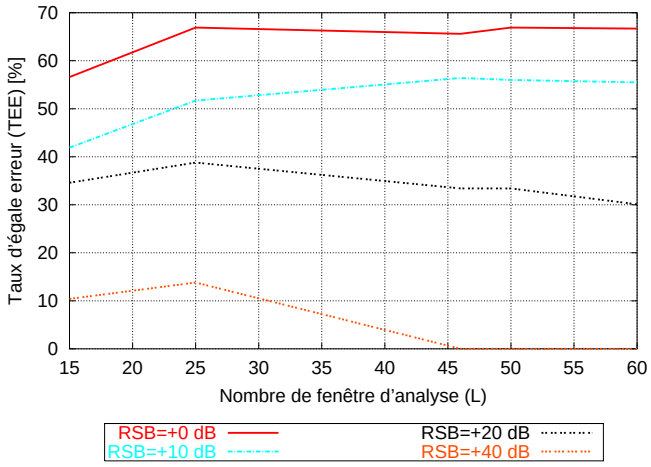


Fig. 3.5: La dépendance des performances de l'algorithme en fonction du nombre L

Un exemple d'application de l'algorithme est montré en Figure 3.6. Le bruit environnemental est un écoulement d'eau et le son à détecter est une sonnerie. Le rapport signal sur bruit moyen est de 0 dB.

La première fenêtre montre l'évolution temporelle du signal analysé. Nous voyons que le son à détecter est bien caché dans le bruit. La deuxième fenêtre représente la variation temporelle de l'énergie et la troisième, l'énergie normalisée. Nous observons que l'énergie du signal présente une crête au moment de l'apparition du signal utile, mais celle-ci reste en deçà de l'énergie maximale. Dans la quatrième fenêtre figure la variance de l'énergie normalisée sur laquelle est faite le seuillage en vue de la détection. Enfin, dans la dernière fenêtre nous trouvons le signal de détection (de type binaire $0 =$ «*pas de détection*» et $1 =$ «*détection*»). Le signal de détection indique une fausse alarme à l'instant t=12s qui correspond à l'apparition du bruit d'écoulement d'eau.

En observant, dans l'exemple ci-dessus, la fenêtre affichant la variance de l'énergie, on remarque que le choix de la valeur du seuil de détection est extrêmement critique, car la différence entre les valeurs constantes et les valeurs à l'instant où il devrait y avoir détection est extrêmement faible. En conclusion, le seuil de l'algorithme semble difficile à régler en conditions réelles.

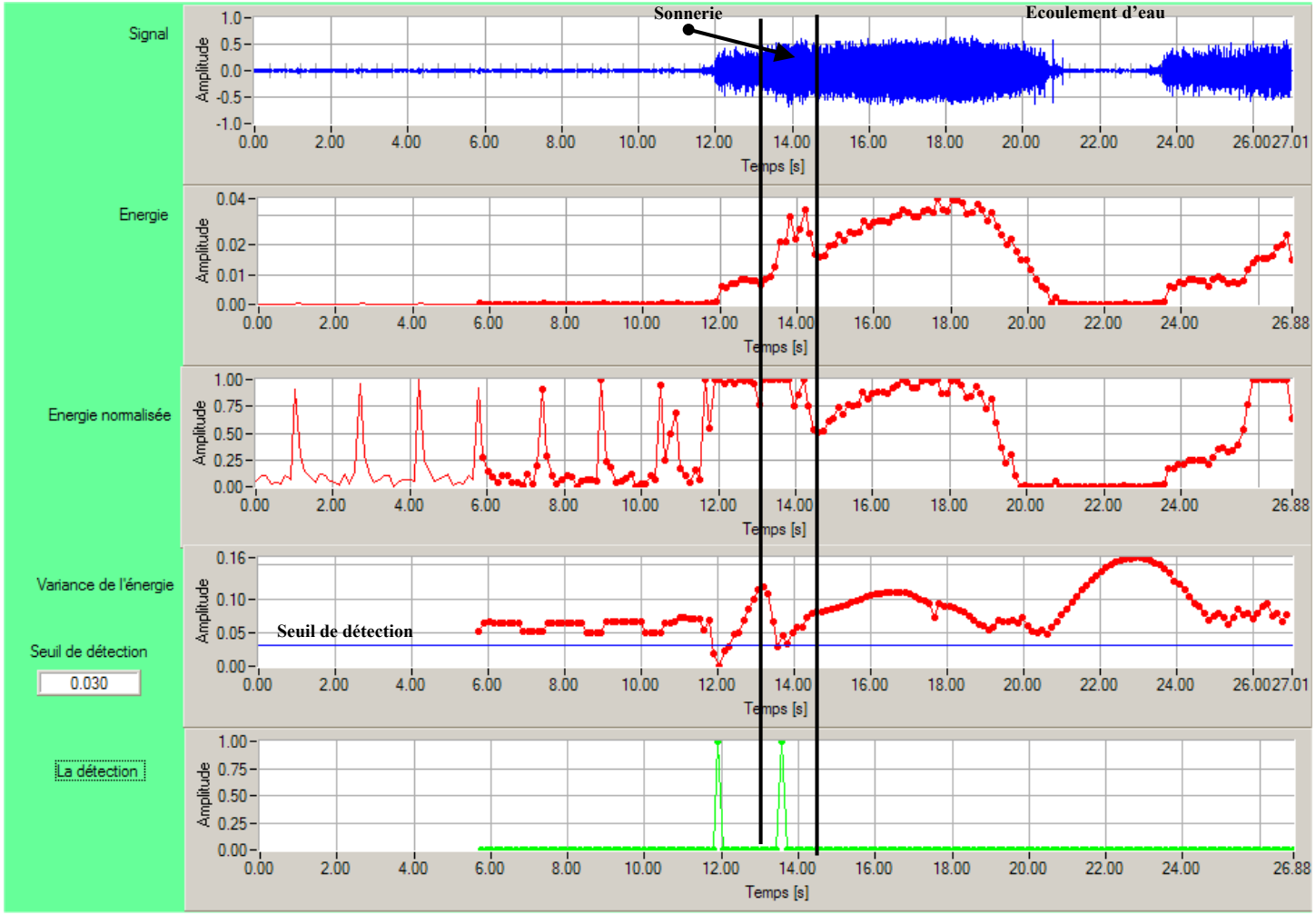


Fig. 3.6: Algorithme fondé sur la variance de l'énergie appliqué à un signal comprenant un signal utile sous forme d'une sonnerie apparaissant à l'instant t=13.5s (RSB moyen de 0 dB), et un bruit de fond correspondant à un écoulement de d'eau

3.2.1.2 Algorithme avec filtre médian

Cet algorithme détecte l'apparition d'un signal en utilisant comme fonction $p(o)$ (voir introduction) la différence entre l'énergie filtrée par un filtre médian conditionné et l'énergie. Pour faciliter la compréhension de l'algorithme, le filtre médian et le filtre médian conditionné sont également décrits dans cette partie.

Le filtre médian Le filtre médian est un type de filtre non-linéaire qui s'applique sur une fenêtre de L échantillons (L est un nombre impair) et il délivre en sortie la valeur correspondant à l'indice $\frac{L-1}{2}+1$ après avoir ordonné les L valeurs dans un ordre croissant ou décroissant [Chang and Wu, 2000], [Tepedelenlioglu et al., 2001]. Son organigramme est donné en Figure 3.7. Ce filtre est non-linéaire et non causal de type passe-bas.

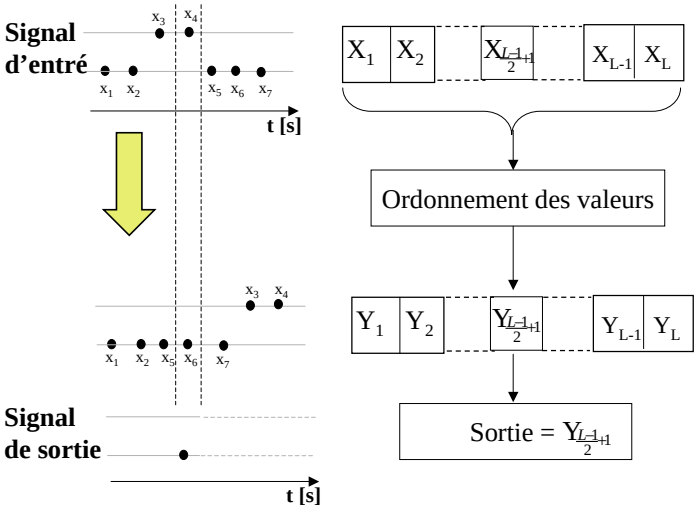


Fig. 3.7: Organigramme du filtre médian

Pour montrer le fonctionnement d'un filtre de longueur $L = 7$ nous avons injecté en entrée une série d'impulsions de largeur variable entre 1 et 10 échantillons (Figure 3.8). On observe que les impulsions ayant une largeur de 1 à 3 échantillons sont éliminées par le filtre médian parce qu'elles ont une largeur plus petite ou égale [2] à $\frac{L-1}{2}$ (la mi largeur du filtre). Nous pouvons conclure que les signaux ayant une largeur plus petite ou égale à $\frac{L-1}{2}$ sont éliminés.
Remarque : Le signal en sortie de filtre médian présente un retard par rapport à l'entrée de $\frac{L-1}{2}$ fois la durée d'un échantillon (effet de la non causalité du filtre).

[2] Toutes les valeurs sont nulles sauf un nombre inférieur ou égal à $\frac{L-1}{2}$, donc la valeur médianne sera nulle

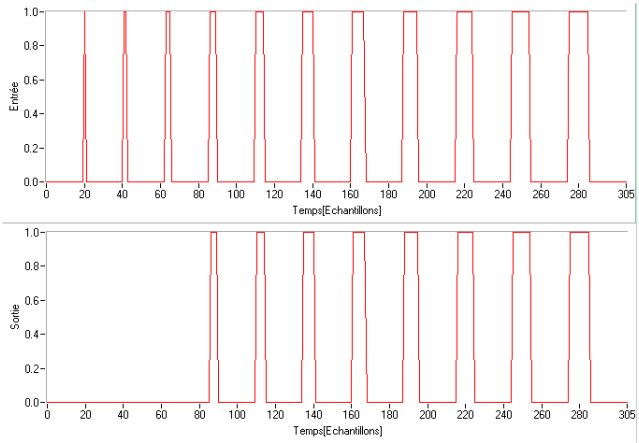


Fig. 3.8: Réponse du filtre médian à une série d'impulsions de largeur variable

Le filtre médian est utilisé dans plusieurs domaines comme dans le traitement d'images pour éliminer le bruit additif aléatoire. Il est utilisé pour l'élimination des impulsions avec une durée plus courte que sa longueur divisé par deux.

Le Filtre médian conditionné Le filtre médian conditionné se comporte comme un filtre médian seulement si la différence entre la valeur de sortie du filtre médian et celle du signal d'entrée dépasse un seuil. Dans le cas contraire le signal de sortie est égal au signal d'entrée. Son organigramme est présenté en Figure 3.10. Ce filtre est non-linéaire et non causal de type passe-bas.

Par exemple, si nous appliquons à l'entrée du filtre médian conditionné une série d'impulsions de largeur de 2 échantillons, les impulsions de largeur de 2 échantillons seront éliminées par un filtre médian simple de largeur 7. L'amplitude de ces impulsions est croissante entre 1 et 10 comme on observe dans la Figure 3.9. Étant donné que la largeur de l'impulsion est de 2 pour une largeur du filtre égale à 7, il y aura une sortie non nulle du filtre, si et seulement si, l'impulsion est située au centre du filtre et si la différence entre l'amplitude de l'impulsion et celle du signal de la sortie du filtre médian associé est inférieure au seuil. Cette fois, les impulsions de largeur inférieure ou égaux à $\frac{L-1}{2}$ sont éliminées, sauf si leur amplitude est inférieure au seuil.

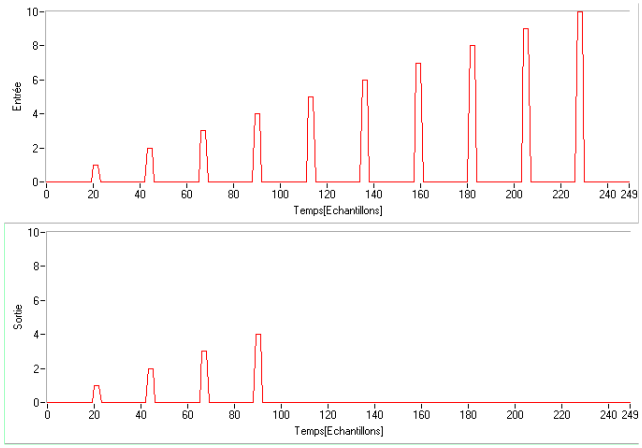


Fig. 3.9: Réponse du filtre médian conditionné ($L = 7$) à une série d'impulsions d'amplitude variable de largeur égale à 2($2 < \frac{L-1}{2}$)

Quand à l'entrée du filtre on applique la même série d'impulsions que précédemment mais avec une largeur de 4 échantillons, on obtient les résultats présentés dans la Figure 3.11. Dans ce cas toutes les impulsions sont transmises, indifféremment du seuil du filtre :

1° si le seuil est dépassé, le filtre médian ne peut pas les éliminer, donc les impulsions sont transmises à cause de leur largeur supérieure à $\frac{L-1}{2}$;

2° dans le cas contraire, le signal est transmis en sortie sans filtrage.

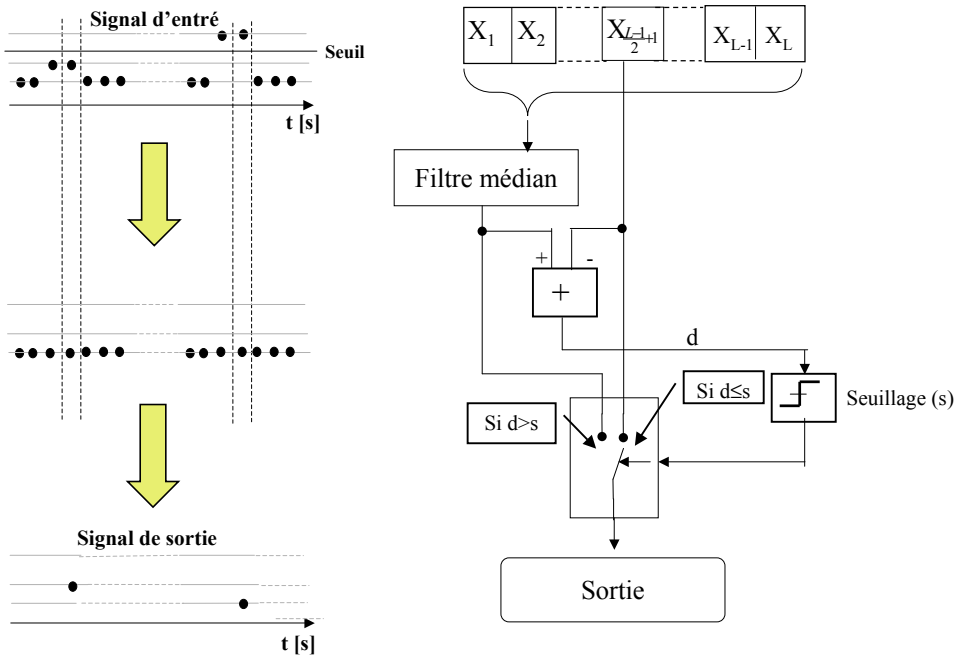


Fig. 3.10: Organigramme du filtre médian conditionné

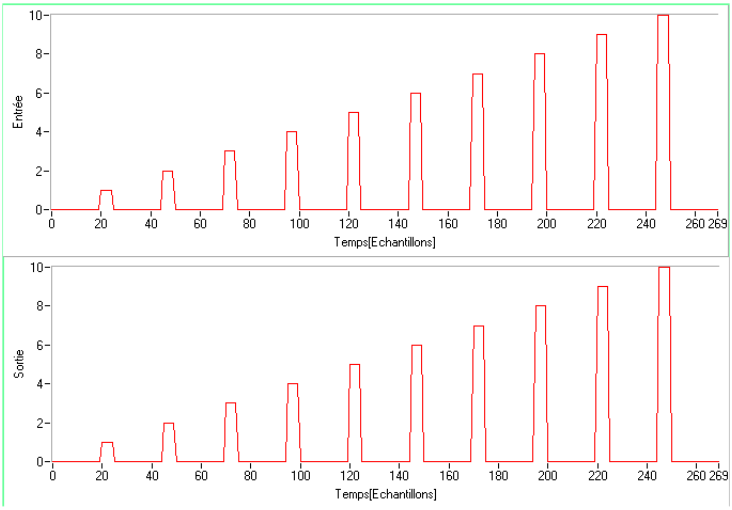


Fig. 3.11: Réponse du filtre médian conditionné ($L = 7$) à une série d'impulsions d'amplitude variable de largeur égale à 4 ($4 > \frac{L-1}{2}$)

L'algorithme de détection proprement dit L'organigramme de cet algorithme est présenté en Figure 3.12. Cet algorithme utilise la propriété du filtre médian conditionné qui permet d'éliminer les variations rapides du signal d'entrée. En calculant la différence entre la sortie du filtre médian conditionné et son entrée on obtient un signal comprenant seulement les variations rapides du signal d'entrée. L'avantage de l'utilisation du filtre médian conditionné par rapport au filtre médian simple est qu'on élimine l'influence des variations de petite amplitude qui peuvent provenir du bruit de fond.

Les étapes de cet algorithme sont :

- Calcul de l'énergie moyenne par échantillon du signal selon formule (3.4).
- Filtrage de l'énergie avec un filtre médian conditionné de longueur $L = 7$. Le seuil utilisé est fixé à 0.001 (voir remarque 2). Le signal résultant est noté $E_{\text{filtré}}$
- Calcul de la différence entre l'énergie et l'énergie filtrée par le filtre médian conditionné. Nous appellerons ce signal «Diff».

$$\text{Diff}(i) = E(i) - E_{\text{filtré}}(i) \tag{3.9}$$

- Seuillage sur le signal de différence «Diff» ; si le seuil (λ) est dépassé, il y a une détection.

$$d(\text{Diff}) = \begin{cases} 1 \text{ si} & \text{Diff} \geq \lambda \\ 0 \text{ si} & \text{Diff} < \lambda \end{cases} \tag{3.10}$$

Remarques :

1. La longueur du filtre médian L dépend de la longueur des impulsions à éliminer et du retard maximal toléré entre le signal d'entrée et celui de la sortie (le retard est égal à $\frac{L-1}{2}$). Nous avons choisi $L = 7$ d'où un retard de $\approx 0.44s$ (la valeur maximale acceptable par notre application est de 0.5 s).

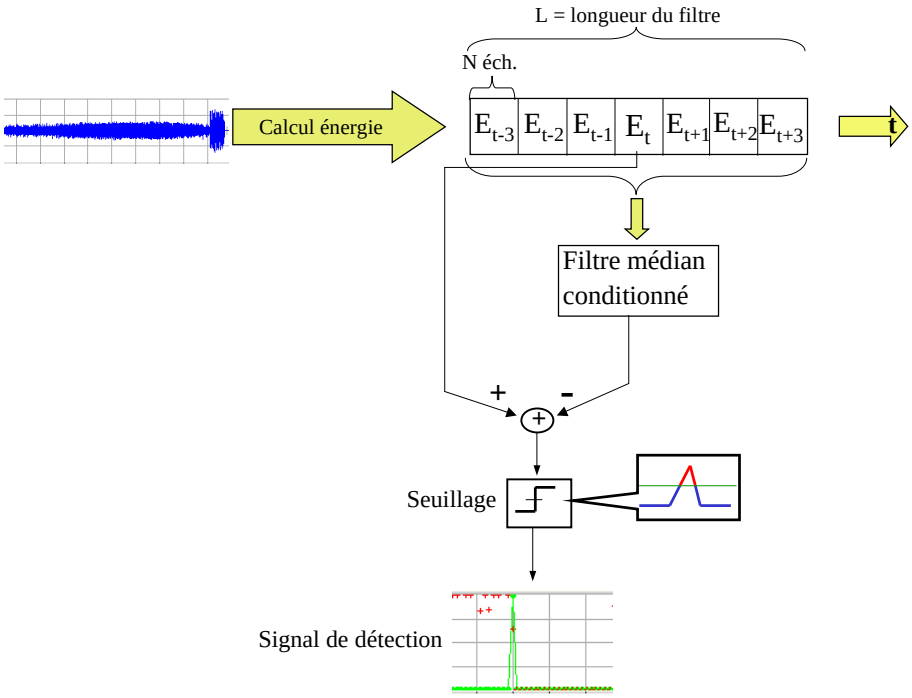


Fig. 3.12: Organigramme de l'algorithme fondé sur le filtre médian conditionné

2. La valeur du seuil du filtre médian conditionné est extrêmement critique pour le bon fonctionnement de l'algorithme. La Figure 3.13 montre la dépendance entre les performances de l'algorithme et le seuil du filtre median conditionné pour notre corpus. Nous avons trouvé une valeur [3] optimale de 0.001 pour notre corpus et le bruit HIS.

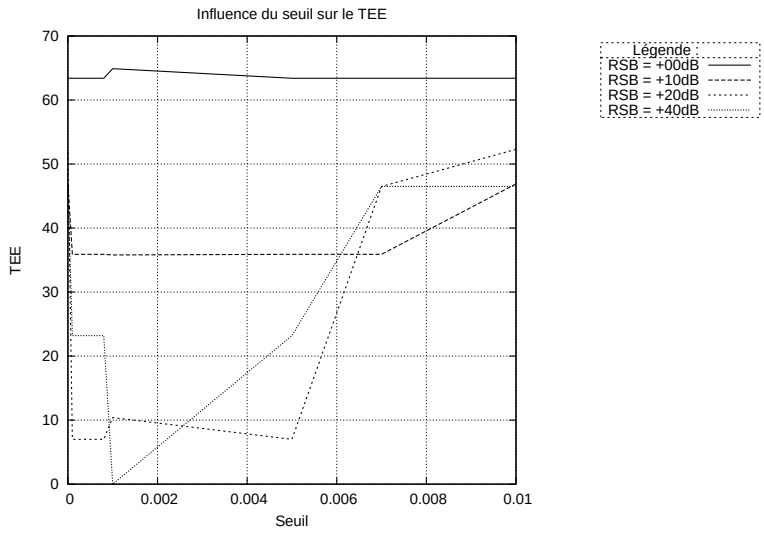


Fig. 3.13: Dépendance des performances de l'algorithme en fonction du seuil du filtre médian conditionné

[3] Il faut savoir que ce seuil est appliqué sur une courbe dont les valeurs varient entre 0 et 1

On trouve un exemple d'application de l'algorithme dans la Figure 3.14 sur le même signal que pour la Figure 3.6.

La première fenêtre montre l'évolution temporelle du signal analysé. Nous voyons que le signal à détecter est bien caché dans le bruit. La deuxième fenêtre visualise la variation temporelle de l'énergie. Nous observons que l'énergie du signal présente une crête au moment de l'apparition du signal utile, mais son amplitude est faible. La fenêtre suivante montre le résultat du filtrage de l'énergie avec le filtre médian conditionné, la quatrième est la différence entre l'énergie du signal et le résultat du filtrage médian conditionné. C'est sur cette différence que se fait le seuillage en vue de la détection. Enfin, la dernière fenêtre comprend le signal de détection (de type binaire $0 =$ «*pas de détection*» et $1 =$ «*détection*»).

Ce type d'algorithme impose de régler, non seulement, le seuil de détection mais aussi celui du filtre médian conditionné (les deux seuils sont indépendants). Cette méthode, mise au point dans le cas du bruit blanc [Dufaux, 2001], sera donc très difficile à mettre en pratique dans des conditions expérimentales réelles, le seuil du filtre dépendant du bruit présent dans l'environnement.

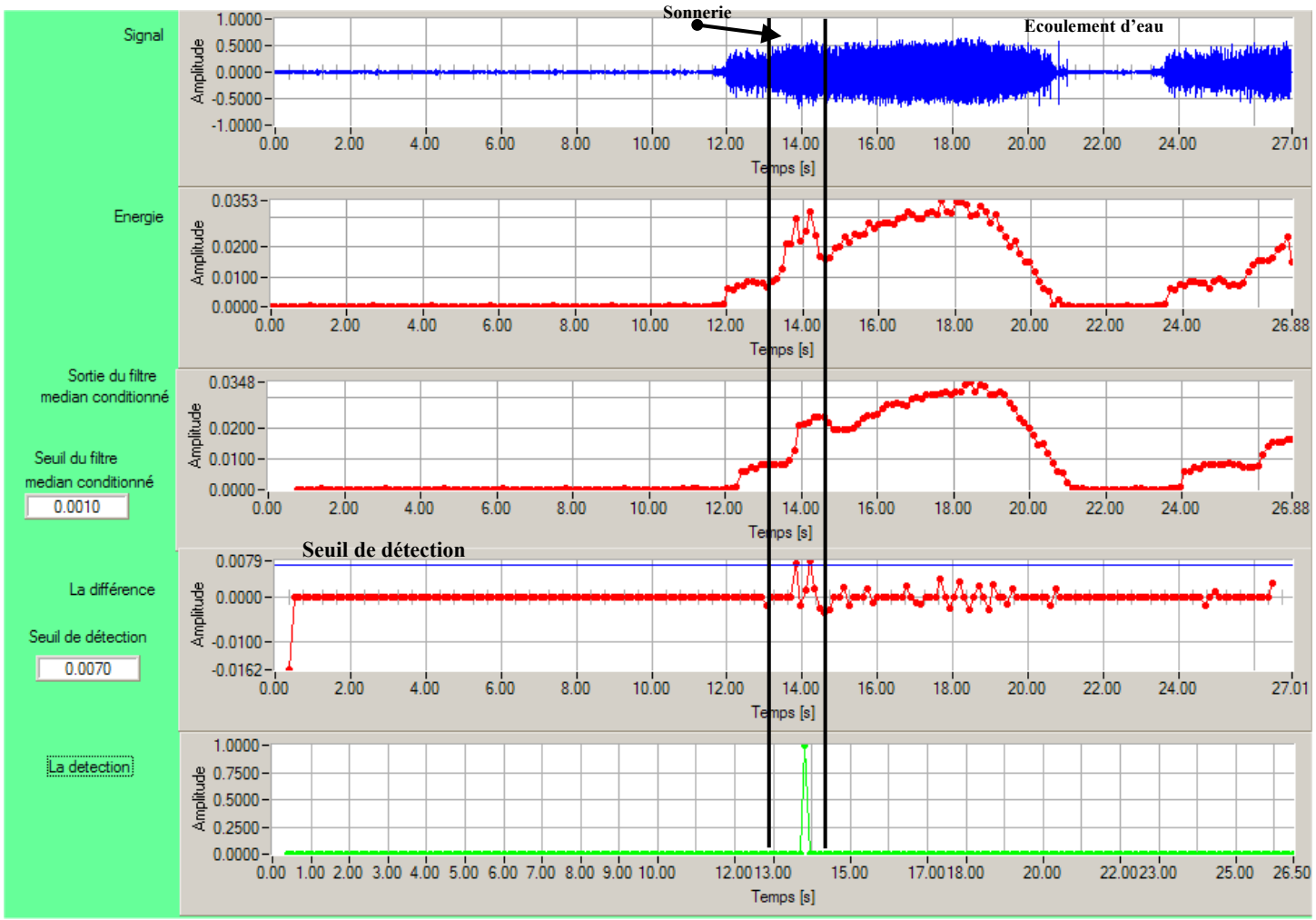


Fig. 3.14: Algorithme fondé sur le filtre médian appliqué à un signal comprenant un signal utile sous forme d'une sonnerie apparaissant à l'instant t=13.5s (RSB moyen de 0 dB), et un bruit de fond correspondant à un écoulement d'eau

3.2.1.3 Algorithme avec filtre médian et seuil adaptatif

Cet algorithme calcule un seuil adaptatif qui est appliqué directement sur l'énergie du signal ($p(o) = E$ - voir introduction), à l'aide d'un filtre médian et d'un filtre médian conditionné. L'organigramme de cet algorithme est présenté dans la Figure 3.15.

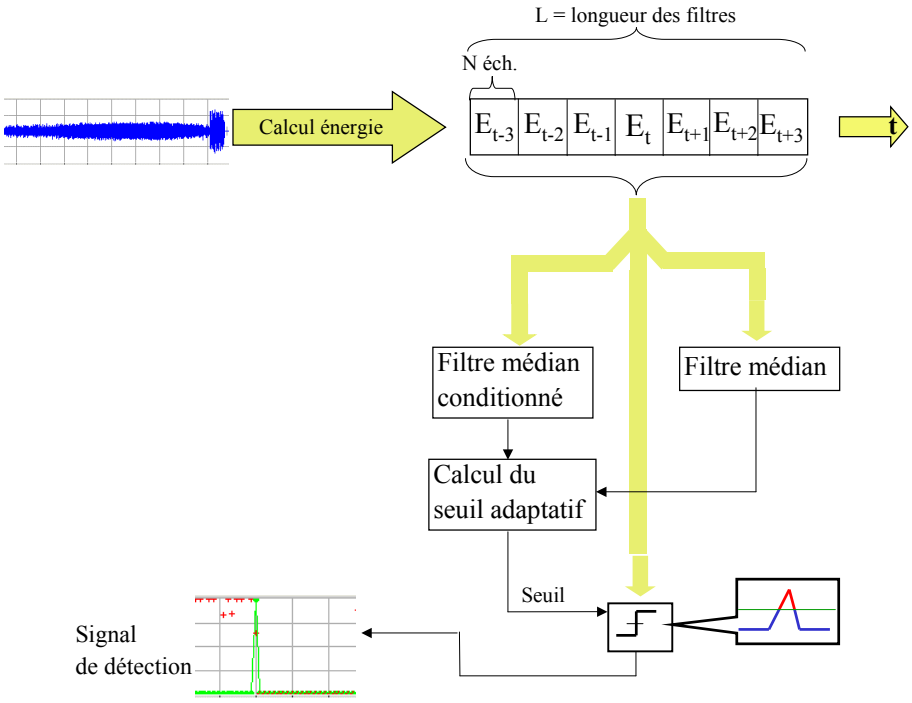


Fig. 3.15: Organigramme de l'algorithme avec seuil adaptatif

Les étapes de cet algorithme sont les suivantes :

- Calcul de l'énergie moyenne par échantillon du signal avec la formule (3.4).
- Filtrage de l'énergie avec un filtre médian de longueur [4] $L = 7$ pour obtenir $E_{\text{filtré médian}}$ et avec un filtre médian conditionné pour obtenir $E_{\text{filtré médian conditionné}}$ avec un seuil [5] de 0.001. Le choix de ce seuil est critique et a été effectué pour l'algorithme précédent.

- Calcul de la valeur du seuil (λ) avec :

$$\lambda(i) = 0.9 \cdot (E_{\text{filtré médian conditionné}}(i) - E_{\text{filtré médian}}(i)) + \kappa \cdot E_{\text{filtré médian}}(i) \tag{3.11}$$

  où κ est une constante.
- Seuillage (3.11) sur l'énergie du signal ; il y a détection si le seuil est dépassé.

$$d(E) = \begin{cases} 1 \text{ si} & E \geq \lambda \\ 0 \text{ si} & E < \lambda \end{cases} \tag{3.12}$$

[4] Voir la Remarque 1 de la page 63
[5] Voir la Remarque 2 de la page 63

Le fonctionnement de l'algorithme est montré dans la Figure 3.16 en utilisant le même fichier de données que pour les algorithmes que nous avons détaillés précédemment.

La première fenêtre visualise l'évolution temporelle du signal analysé. La deuxième fenêtre montre la variation temporelle de l'énergie et la variation du seuil adaptatif qui sera utilisé dans la détection. Dans les deux fenêtres qui suivent sont présentés les résultats du filtrage de l'énergie avec un filtre médian et avec un filtre médian conditionné. La valeur du seuil est obtenu au moyen de la formule (3.11). Nous pouvons observer que le seuil suit la variation de l'énergie du bruit de fond et permet la détection du signal utile. Enfin, dans la dernière fenêtre nous avons visualisé le signal de détection (de type binaire $0 =$ «*pas de détection*» et $1 =$ «*détection*»). Le signal de détection présente 4 fausse alarmes.

Remarque : Les signaux de sortie des filtres médian et médian conditionné, et implicitement celui de la détection commencent à être calculés au bout de 0.8s seulement, parce que l'algorithme a besoin d'un temps d'initialisation.

Nous pouvons faire les mêmes remarques que dans le cas de l'algorithme précédent (l'algorithme fondé sur le filtre médian) sur le fait qu'il y a deux seuils indépendants à régler. En particulier, le seuil du filtre médian conditionné sera très difficile à optimiser en fonction d'un bruit expérimental qui n'a pas les propriétés d'un bruit blanc.

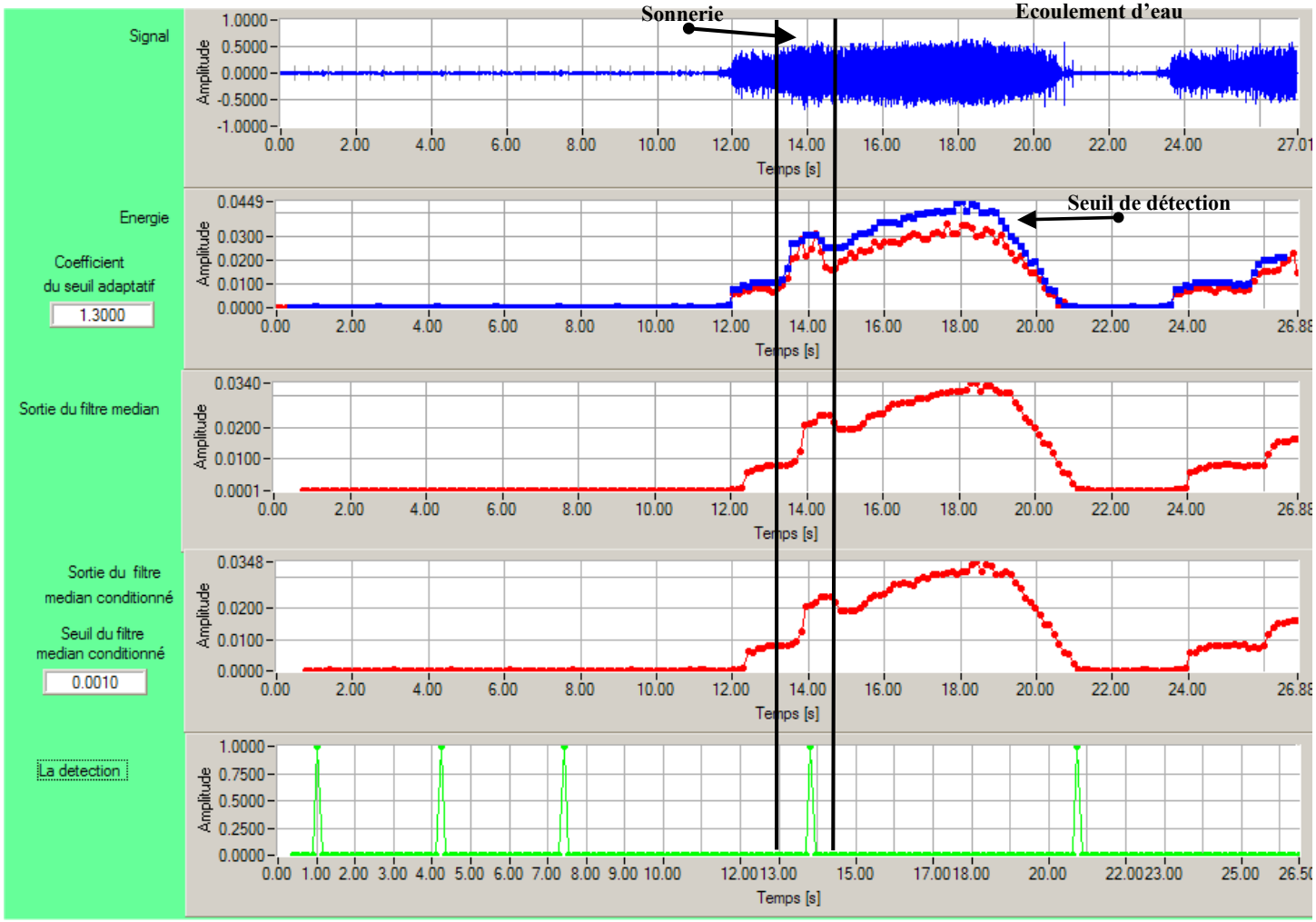


Fig. 3.16: Algorithme avec seuil adaptatif appliqué à un signal comprenant un signal utile sous forme d'une sonnerie apparaissant à l'instant t=13.5s (RSB moyen de 0 dB), et un bruit de fond correspondant à un écoulement d'eau

3.2.1.4 Évaluation de ces trois algorithmes issus de l'état de l'art sur notre corpus de sons (DSIM)

Les trois algorithmes présentés dans les sections précédentes 3.2.1.1 - 3.2.1.3 ont été testés sur le corpus de sons pour la détection (voir section 2.2).

Méthodologie d'analyse des performances Pour comparer les performances des algorithmes de détection présentés, nous avons calculé le taux de détections manquées (TDM) selon (3.13) et le taux de fausses alarmes (TFA) selon (3.14). Le taux de détections manquées est le rapport entre le nombre des signaux non-détectés par l'algorithme et le nombre total des signaux à détecter. Le taux de fausses alarmes est le rapport entre le nombre des détections qui ne correspondent pas à un signal et la somme du même nombre avec le nombre total des signaux à détecter. Le nombre de fausses alarmes est rajouté au dénominateur pour s'assurer que la valeur du taux se situe dans l'intervalle $[0, 1]$.

$$TDM = \frac{\text{Le nombre de détections manquées}}{\text{Le nombre de signaux à détecter}} \tag{3.13}$$

$$TFA = \frac{\text{Le nombre de fausses alarmes}}{\text{Le nombre de fausses alarmes} + \text{Le nombre de signaux à détecter}} \tag{3.14}$$

Nous considérons comme une *fausse détection*, une détection qui intervient plus de 0.5s avant le moment du début du signal à détecter ou qui survient après la fin du signal à détecter. Le choix de cet intervalle de temps situé avant le début du signal est dû à la non causalité des filtres de type médian : il a été fait en considérant le cas le plus défavorable dans lequel une petite partie du signal à détecter survient à la fin de la dernière fenêtre d'analyse. Étant donné que pour les deux algorithmes fondés sur le filtre médian, l'analyse du signal est faite sur 4 fenêtres successives contenant 2048 échantillons, il résulte un intervalle d'erreur possible de largeur $\approx$0.5s. Pour l'algorithme fondé sur la variance de l'énergie, l'analyse du signal se fait sur une fenêtre unique de 2048 échantillons qui conduit à une erreur possible de $\approx$0.13s. En conséquence, une erreur de 0.5 s est admise pour tous les algorithmes dans le cadre de notre méthodologie d'évaluation.

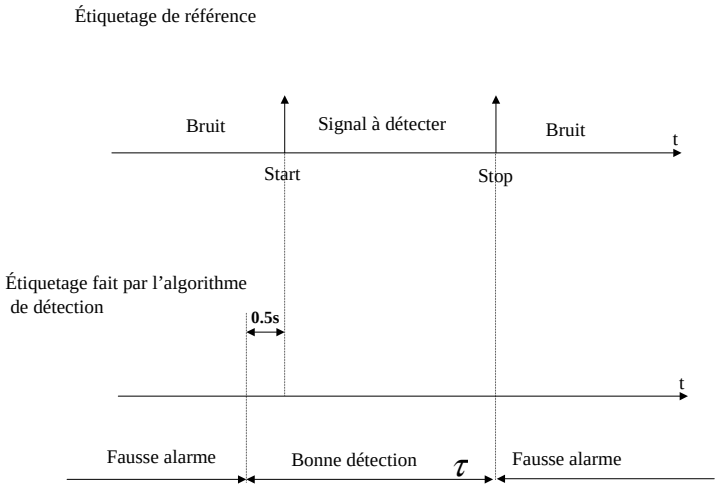


Fig. 3.17: Définition et localisation des fausses alarmes et des bonnes détections

Une détection est considérée comme *bonne* lorsqu'elle se situe dans l'intervalle de temps (τ) débutant 0.5s avant le signal et se terminant à la fin du signal. Une détection est considérée comme *manquée* si aucune détection n'est observée dans l'intervalle de temps τ défini précédemment, alors qu'une détection aurait dû être faite. Ces définitions sont illustrées dans la Figure 3.17.

Quand nous avons généré les fichiers de tests nous avons aussi généré pour chacun un fichier d'étiquetage (de type «.sam») avec l'étiquette du début du son et de sa fin qui permet le calcul des performances des algorithmes. Un exemple de fichier type «.sam» est inséré dans l'annexe D.1. Pour plus de détails, il convient de se reporter à la norme «sam» pour les corpus de parole qui est explicitée dans [Standard SAM, 1992], [Wells et al., 1992].

Nous avons donc tracé pour chaque algorithme la courbe du taux de détections manqués en fonction du seuil de décision, la courbe du taux de fausses alarmes en fonction du seuil de décision ainsi que la courbe du taux de détections manquées en fonction du taux de fausses alarmes ; c'est à partir de cette dernière que nous avons pu déduire le taux d'égale erreur (TEE) qui est défini par l'équation (3.15). Ces courbes ont été tracées séparément pour chaque type de bruit (bruit blanc, bruit de HIS et bruit de l'eau) et globalement.

$$TEE = TDM\Big|_{TDM=TFA} \tag{3.15}$$

Performances Les performances obtenues pour les trois algorithmes appliqués sur le corpus de sons (DSIM) pour la détection sont présentées dans le tableau 3.2. Dans la première colonne se trouve le nom de l'algorithme, dans la deuxième le rapport signal sur bruit et dans les trois colonnes suivantes, les valeurs du taux d'égale erreur pour différents types de bruit de fond (bruit blanc, bruit HIS - le bruit de l'environnement de l'appartement d'étude et bruit d'eau). Les courbes ROC[6] (taux de détections manquées en fonction du taux de fausses détections) sont données dans l'annexe F.1 (à la page 167).

[6] Receiver Operating Characteristic

Pour comparer les algorithmes on va considérer prioritairement le bruit HIS et celui de l'écoulement d'eau (le bruit blanc n'est pas réaliste pour notre application) et comme RSB les valeurs plus petites ou égale à 10 dB (valeurs proches des conditions réelles dans 55% des cas - voir section 2.2.3 de la page 44).

L'algorithme fondé sur la variance de l'énergie, montre de très bonnes performances sur le bruit blanc mais des mauvaises sur le bruit HIS. Dans le cas du bruit de l'eau l'algorithme présente de bonnes performances à l'exception du cas RSB=0 dB. Sur sa courbe ROC (Figure F.1 de la page 167) on observe qu'on obtient un minimum du TDM pour une valeur assez élevée du TFA 0.88, cet algorithme n'est donc pas très bien adapté dans notre cas (pour le bruit HIS). Cet algorithme se fonde sur la normalisation de l'énergie : l'arrivée d'une variation de l'énergie rapide et grande implique la diminution de l'énergie normalisée presque à zéro (normalisation par une grande valeur) pour une courte période. Ce phénomène entraîne la diminution de la variance. Pour le bruit blanc, tant qu'il n'y a pas de son à détecter, l'énergie présente des variations réduites ; de même pour le bruit d'écoulement d'eau qui est un bruit coloré ; par contre pour le bruit HIS, l'énergie a des variations même en l'absence de son. Cela explique les mauvais résultats obtenus dans le cas du bruit HIS.

L'algorithme fondé sur le filtrage médian montre une réponse assez bonne pour le bruit blanc et le bruit d'eau mais toujours non satisfaisante pour le bruit HIS. Sur sa courbe ROC (Figure F.3 de la page 167) on observe que le point qui donne un minimum de TDM pour un RSB de 0 et 10 dB n'est pas un minimum pour 20 et 40 dB. Cet algorithme est difficile à utiliser parce qu'il a deux seuils à régler et celui du filtre médian conditionné dépend beaucoup du signal. Pour le bruit blanc les performances sont assez bonnes parce que le filtrage médian conditionné arrive à éliminer complètement le bruit mais dans le cas du bruit HIS qui présente beaucoup d'impulsions c'est plus difficile (le choix du seuil de détection est aussi difficile). L'énergie du bruit de l'écoulement de l'eau a une variation lente qui permet au filtre médian conditionné de l'éliminer pour des RSB élevés (RSB>0 dB).

L'algorithme qui utilise le filtrage médian pour calculer un seuil de détection adaptatif présente des moindres performances pour les bruits réels (bruit HIS et l'eau) alors qu'il est satisfaisant pour le bruit blanc. Sur la courbe ROC de cet algorithme (Figure F.4 de la page 167) on observe que, de même que pour les deux autres algorithmes, la valeur minimale de TDM est obtenue pour 90% de TFA (ce qui n'est pas acceptable). L'utilisation d'un seuil adaptatif dans le cas des bruits réels n'implique pas l'amélioration des performances de l'algorithme. Cela s'explique par le fait que l'adaptation du filtre à été optimisée pour le cas des bruits blancs ou pseudo-blancs.

Méthode de détection	RSB [dB]	TEE pour les différents types de bruit de fond		
		Bruit blanc [%]	Bruit du HIS [%]	Bruit d'eau [%]
Seuillage de la variance de l'énergie	0	5.7	69.6	41.9
	+10	**0**	**56.4**	**0.9**
	+20	0	33.4	0
	+40	0	5.2	6.1
Seuillage de l'énergie filtrée avec un filtre médian conditionné	0	30	64.9	24.1
	+10	**0**	**35.8**	**6.1**
	+20	0	10.4	0
	+40	0	0	0
Seuillage adaptatif de l'énergie (dépend d'un filtre médian)	0	11.4	64.2	31.5
	+10	**0**	**38.9**	**21.7**
	+20	0	31.8	1
	+40	0	15.7	0

Tab. 3.2: Résultats obtenus avec les trois algorithmes de détection

3.3 Algorithmes de détection proposés

Dans la section précédente l'évaluation des trois algorithmes issus de l'état de l'art a été donnée. Pour le cas du bruit blanc les trois algorithmes ont un taux d'égale erreur de 0% pour RSB$\geq$10dB et de 6%, 30% et respectivement 11% pour RSB=0 dB. En opposition pour le bruit HIS (bruit environnemental de l'appartement d'étude) et pour le bruit d'écoulement d'eau les performances dans le cas des RSB$\leq$20 dB sont inacceptables. Pour ces raisons, cette section propose trois nouveaux algorithmes, mieux adaptés à ce type de bruit.

3.3.1 Algorithme fondé sur la fonction d'intercorrélation

Cet algorithme utilise comme fonction $p(o)$ (voir introduction) le maximum de la fonction d'intercorrélation de deux fenêtres successives du signal. L'algorithme se fonde sur le fait que si le signal analysé correspond à un bruit stationnaire, les 2 fenêtres temporelles successives sont fortement corrélées (valeur importante du maximum de la fonction d'intercorrélation), sauf pour le bruit blanc ou pseudo-blanc ; par contre, si dans la dernière fenêtre un nouveau signal vient se superposer au bruit de fond, ce ne sera plus le cas et on observera une diminution du maximum de la fonction d'intercorrélation. Le calcul du maximum de la fonction d'intercorrélation entre 2 fenêtres temporelles successives doit donc permettre de détecter l'apparition d'un son impulsionnel sur un bruit suffisamment corrélé.

L'organigramme de l'algorithme correspondant est présenté dans la Figure 3.18. Cet algorithme fait un traitement sur 2 fenêtres consécutives de signal. Nous procédons tout d'abord à la normalisation du signal dans chacune des deux fenêtres d'analyse pour éliminer l'influence de l'énergie du signal sur la fonction d'intercorrélation. Les normalisations possibles inclut : normalisation par la valeur maximale du signal dans la fenêtre ou par la racine carrée de l'énergie du signal dans la fenêtre d'analyse. La normalisation par la racine carrée de l'énergie a été choisie pour éliminer l'influence de la variation de l'énergie du signal. L'étape suivante consiste

à calculer la fonction d'intercorrélation entre les signaux normalisés des deux fenêtres consécutives ; ensuite un seuillage sur le signal des valeurs maximales de la fonction d'intercorrélation permet d'établir si un son a été détecté ou non (si cette grandeur descend au-dessous du seuil, il y a une discontinuité dans le signal mesuré et donc détection).

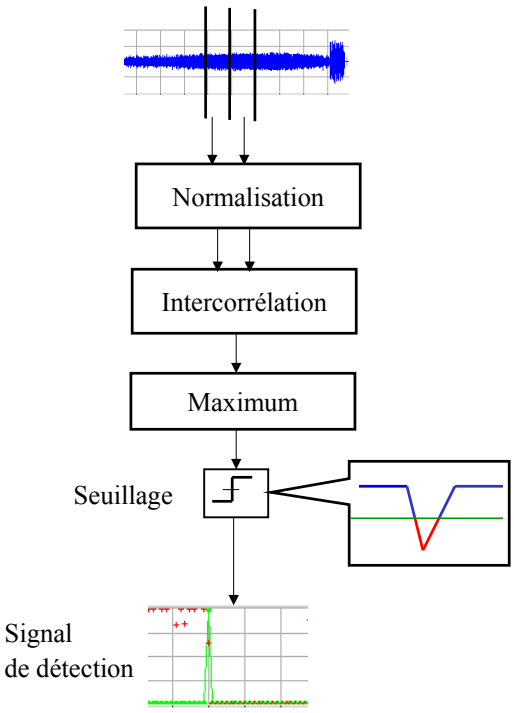


Fig. 3.18: Organigramme de l'algorithme fondé sur la fonction d'intercorrélation

La Figure 3.19 visualise le fonctionnement de l'algorithme pour un claquement de porte comme signal utile et le bruit HIS comme bruit. La première fenêtre visualise l'évolution temporelle du signal analysé. Le signal à détecter ne ressort pas du bruit de manière significative. La deuxième fenêtre montre l'évolution temporelle du maximum de la fonction d'intercorrélation. La différence entre l'amplitude au moment de l'apparition du signal et la valeur de palier est notable. Enfin, dans la dernière fenêtre, nous trouvons le signal de détection (de type binaire $0 =$ «*pas de détection*» et $1 =$ «*détection*»).

Par rapport aux méthodes étudiées précédemment, cette méthode présente l'avantage de n'avoir qu'un seul seuil à régler de surcroît sur un signal normalisé. La mise en oeuvre doit donc en être facilitée.

La fonction d'intercorrélation entre les deux fenêtres consécutives peut être calculée avec la formule de définition (équation (3.16) où $x(k)$ et $y(k)$ sont les deux signaux à corréler, comprenant chacune N échantillons et $\gamma_{xy}(i)$ est la fonction d'intercorrélation discrète). Cette méthode a le désavantage d'être très longue à calculer. La solution le plus souvent utilisée est basée sur la Transformée de Fourier Rapide (TFR) en passant par l'espace des fréquences [Bellanger, 2002]. Dans ce cas le calcul est plus rapide, et l'on utilise le théorème de Wiener-Kintchine (équation (3.17)) où $X(\nu)$ et $Y(\nu)$ sont les transformées de Fourier des signaux et Γ_{XY} est la Densité

Spectrale de Puissance [Kunt et al., 1991].

$$\gamma_{xy}(i) = \sum_{k=-N}^{N} x(k).y(k-i) \tag{3.16}$$

$$\gamma_{XY}(i) = TFR^{-1}[\Gamma_{XY}(\nu)] \quad \text{où} \quad \Gamma_{XY}(\nu) = X(\nu) \cdot Y^*(\nu) \tag{3.17}$$

où TFR^{-1} est la TFR inverse et $Y^*(\nu)$ est la conjuguée complexe de $Y(\nu)$

La taille de la fenêtre d'analyse doit être adaptée aux phénomènes à identifier par intercorrélation et à la précision de mesure du moment de l'apparition de l'événement sonore. La taille de la fenêtre d'analyse doit d'une part être suffisament grande pour que la fonction d'intercorrélation soit représentative des propriétés statistiques du bruit environnemental et elle doit d'autre part être suffisament petite pour disposer d'une bonne précision du moment de détection. La précision de mesure doit être de maximum 0.5s et les impulsions du bruit environnemental ont une durée moyenne de 0.1s. Pour ces raisons, la taille de la fenêtre a été fixée à 256ms (4096 échantillons) avec un chevauchement des fenêtres d'analyse de 50%.

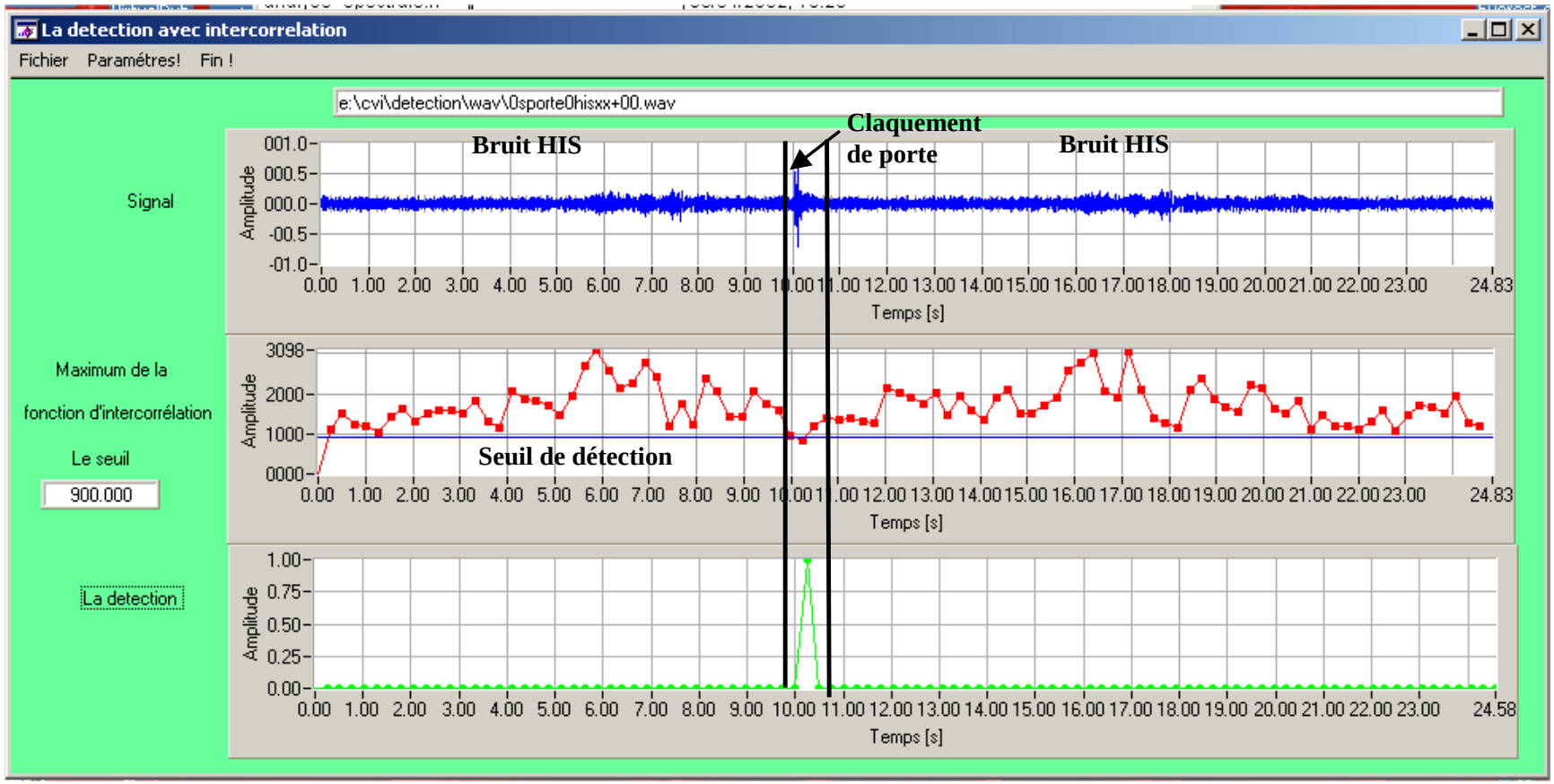


Fig. 3.19: Algorithme fondé sur la fonction d'intercorrélation appliqué à un signal comprenant un signal utile sous forme d'un claquement de porte apparaissant à l'instant t=10s (RSB moyen de 0 dB), et un bruit de fond correspondant au bruit HIS

3.3.2 Algorithme fondé sur la prédiction de l'énergie

Cet algorithme effectue la comparaison entre l'énergie mesurée d'un bloc de $N = 2048$ échantillons et le résultat d'une prédiction faite à partir de la séquence de 10 valeurs successives de l'énergie de blocs de N échantillons mesurés précédemment. Si la valeur absolue de l'écart entre la mesure directe et la prédiction dépasse un certain seuil, cela sera une indication qu'il y a eu apparition d'un nouveau signal sonore que nous aurons donc ainsi détecté.

3.3.2.1 Les fonctions de prédiction SPLINE

Avant de présenter l'algorithme de détection proposé, un petit rappel sur les fonctions d'extrapolation de type SPLINE est fait [Boor, 1978]. Le nom SPLINE en anglais désigne un instrument utilisé par les dessinateurs pour le dessin des courbes qui passe par des points donnés (il est composé d'une règle avec des poids distribués). Pour une fonction donnée $y_i = y(x_j)$, $j = 1 \ldots N$, la fonction d'extrapolation des SPLINE cubiques sur l'intervalle $[x_j, x_{j+1}]$ a la forme générale de l'équation (3.18). Le but des SPLINE est d'obtenir une formule d'extrapolation avec la première et la deuxième dérivée continues sur l'intervalle d'extrapolation.

$$y = Ay_j + By_{j+1} + Cy_j^{''} + Dy_{j+1}^{''} \tag{3.18}$$

où

$$A = \frac{x_{j+1} - x}{x_{j+1} - x_j} \quad B = 1 - A = \frac{x - x_j}{x_{j+1} - x_j}$$

$$C = \frac{1}{6}(A^3 - A)(x_{j+1} - x_j)^2 \quad D = \frac{1}{6}(B^3 - B)(x_{j+1} - x_j)^2$$

Le système d'équations qui permet le calcul de la deuxième dérivée est (3.19), qui est un système de N-2 équations avec N inconnus. Pour avoir une solution unique on impose $y_1^{''}$ et $y_N^{''}$ égales à zéro et nous obtenons les fonctions SPLINE naturelles.

$$\frac{x_j - x_{j-1}}{6}y_{j-1}^{''} + \frac{x_{j+1} - x_{j-1}}{3}y_j^{''} + \frac{x_{j+1} - x_j}{6}y_{j+1}^{''} = \frac{y_{j+1} - y_j}{x_{j+1} - x_j} - \frac{y_j - y_{j-1}}{x_j - x_{j-1}} \tag{3.19}$$

L'avantage des fonctions d'extrapolation cubiques est que le système (3.19) est non seulement linéaire mais qu'il est caractérisé par une matrice tridiagonale qui en simplifie la résolution numérique [Press et al., 2002].

3.3.2.2 L'algorithme proprement dit

L'organigramme de l'algorithme fondé sur la prédiction de l'énergie se trouve en Figure 3.20.

Les calculs sont organisés de la façon suivante :

- acquisition de 2048 échantillons sur le fichier de test et calcul de l'énergie de ce bloc (avec équation 3.4)
- calcul de la valeur absolue de l'écart (ε) entre cette valeur expérimentale de l'énergie et la valeur extrapolée à partir des L valeurs précédentes et calcul de la valeur moyenne (μ) et de l'écart-type (σ) de l'énergie sur les L valeurs :

$$\varepsilon = |E_{\text{extrapolée}} - E| \tag{3.20}$$

$$\sigma = \sqrt{\frac{1}{L-1} \cdot \sum_{i=0}^{L-1} (\varepsilon(i) - \mu)^2} \qquad \text{où} \qquad \mu = \frac{1}{L} \cdot \sum_{i=0}^{L-1} \varepsilon(i) \tag{3.21}$$

- détermination du seuil adaptatif (λ) et obtention du signal de détection par seuillage (κ est une constante)

$$\lambda = \kappa + \sqrt{2} \cdot \sigma + \mu \tag{3.22}$$

$$d(\varepsilon) = \begin{cases} 1 \text{ si} & \varepsilon \geq \lambda \\ 0 \text{ si} & \varepsilon < \lambda \end{cases} \tag{3.23}$$

- décalage vers la droite du buffer de L valeurs et insertion de la dernière valeur à gauche
- retour à la première étape .

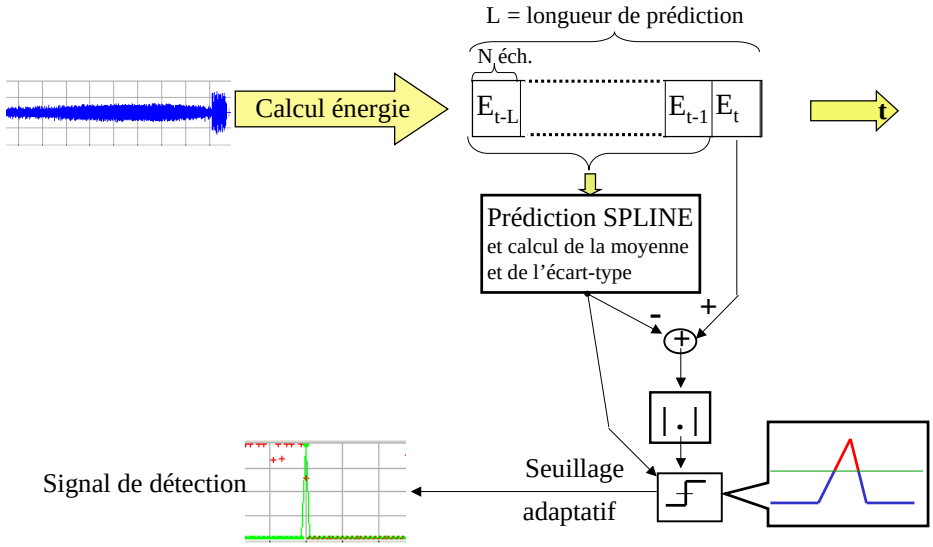


Fig. 3.20: Organigramme de l'algorithme fondé sur la prédiction de l'énergie

On trouve un exemple d'application de l'algorithme dans la Figure 3.21 sur le fichier audio précédemment utilisé pour illustrer les autres algorithmes. La première fenêtre montre l'évolution temporelle du signal analysé. Nous voyons que le signal à détecter est bien caché dans le bruit. Les deux fenêtres suivantes présentent l'énergie du signal ainsi que l'écart de prédiction. Nous pouvons observer que cet algorithme élimine bien les fluctuations d'un signal stationnaire comme celui de l'écoulement de l'eau et met en évidence la sonnerie de téléphone. Enfin, la dernière fenêtre montre l'évolution du signal de détection (de type binaire $0 =$ «*pas de détection*» et $1 =$ «*détection*»).

Le seuil adaptatif est calculé avec la formule (3.22) où la valeur $\sqrt{2}$ permet d'obtenir à partir de l'écart-type une grandeur qui s'apparente à la valeur crête de l'ondulation et la moyenne μ permet de suivre le signal. La valeur du terme κ varie pour obtenir différents points de la courbe ROC pour cet algorithme.

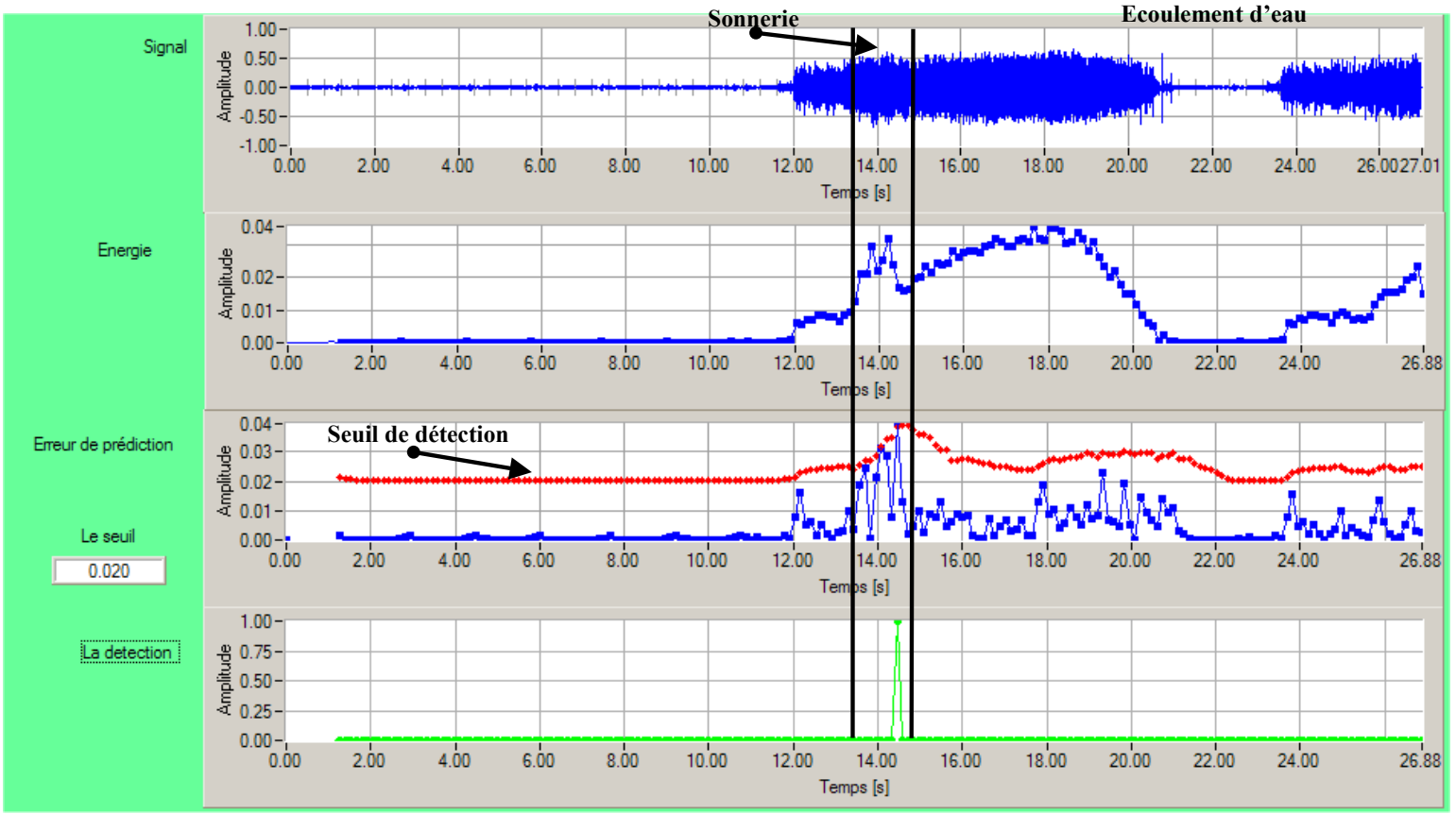


Fig. 3.21: Algorithme fondé sur la prédiction de l'énergie appliqué à un signal comprenant un signal utile sous forme d'une sonnerie apparaissant à l'instant t=13.5s (RSB moyen de 0 dB), et un bruit de fond correspondant à un écoulement d'eau

Le seuillage étant réalisé sur un signal non normalisé, le paramètre κ doit être réglé lors de l'initialisation du système, par enregistrement, en imposant qu'aucun signal utile ne soit présent par exemple pendant les premières 10 secondes. La mise en oeuvre serait donc légèrement plus complexe que celle de l'algorithme fondé sur la fonction d'intercorrélation.

3.3.3 Algorithme fondé sur la décomposition en ondelettes

3.3.3.1 La transformée en ondelettes

Avant de présenter l'algorithme, un rappel est donnée sur les transformées temps-fréquence et plus particulièrement sur la transformée en ondelettes.

Les signaux stationnaires se décomposent en combinaison linéaire d'ondes ce qui fait de la transformée de Fourier l'outil adapté à l'étude de ces signaux. Pour traiter les signaux impulsionnels on segmente l'onde et on la découpe en un morceau d'onde, c'est-à-dire une ondelette, en introduisant un début et une fin. Ces signaux se décomposent de la même façon que les signaux stationnaires, en une combinaison linéaire d'ondelettes.

Tout signal $x(t)$ peut se décomposer en somme d'ondelettes $\psi_{u,s}(t)$ localisées en temps et en fréquence, pondérées par des coefficients $\kappa_{u,s}$:

$$x(t) = \sum_{u,s} \kappa_{u,s} \psi_{u,s}(t) \tag{3.24}$$

La forme des $\psi_{u,s}(t)$ distingue les ondelettes «temps-fréquence» (la Transformée de Fourier à court terme) des ondelettes «temps-échelle» (la transformée en ondelettes).

Les ondelettes «temps-fréquence» sont représentées par la transformée de Fourier à court terme ou la transformée de Gabor. Dans ce cas, l'ondelette $\psi_{u,s}(t)$ est le produit entre une onde et une enveloppe : $\psi_{u,s}(t) = e^{2i\pi sft} b(t - sl)$. L'onde $e^{2i\pi sft}$ localise l'ondelette à la fréquence kf et l'enveloppe $b(t - sl)$ localise ondelette au temps sl. L'atome temps-fréquence occupé est un rectangle symbolique (voir figure 3.22) autour de (sl, kf) avec la surface de $\frac{1}{4\pi}$. L'encombrement est minimal selon le principe d'incertitude de Gabor-Heisenberg. La valeur l est la longueur temporelle occupée par l'ondelette. Dans le cas de la transformée de Fourier à court terme ou de la transformée de Gabor, le signal est segmenté temporellement en fenêtres régulières. Le seul élément modifiable est la longueur l des fenêtres de l'analyse fréquentielle locale ce qui modifie la durée de l'enveloppe $b(t)$. Dans la figure 3.23 (b) le pavage typique d'une transformée de Fourier à court terme est montré. Les cas extrêmes (a) et (c) de la même figure permettent de partager le plan temps-fréquence en base de Dirac et base de Fourier. Cependant, le pavage ne peut pas être modifié, le partitionnement étant rigide. Le pavage n'a qu'un seul degré de liberté qui est la longueur des pavés.

Les ondelettes «temps-échelle» sont obtenues avec la transformée en ondelettes (DWT). Ces ondelettes sont définies dans le cas général à partir d'une ondelette mère qui subit un changement d'échelle de facteur s et qui est décalée en temps de u :

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right), \tag{3.25}$$

où $u = sjl$ et s a la forme $s = \alpha^k$.

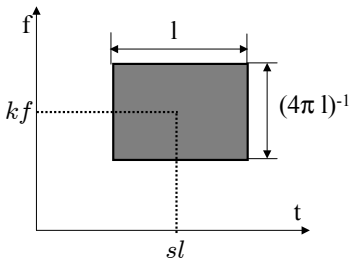


Fig. 3.22: Localisation et encombrement spectral d'une cellule élémentaire pour la transformée de Fourier à court terme

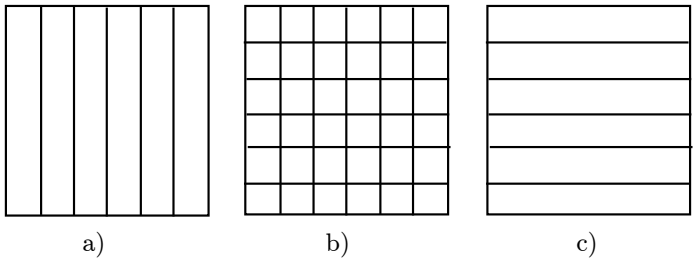


Fig. 3.23: Plan temps-fréquence partitionné en : a) Base de Dirac, b) Base de Fourier fenêtrée, c) Base de Fourier

Le paramètre α de compression/dilatation localise l'ondelette à la fréquence $\frac{f_0}{\alpha^k}$ (f_0 est la fréquence centrale de l'ondelette mère) et à l'instant $\alpha^k jl$. Pour modifier le pavage, nous pouvons changer le nombre d'échelles possibles, le nombre de voies (voir figure 3.24) ainsi que l'ondelette mère qui définit l'encombrement temps-fréquence (voir figure 3.25). L'allure du partitionnement reste la même, avec une bonne résolution temporelle pour les hautes fréquences et une moins bonnes résolution temporelle en basse fréquence.

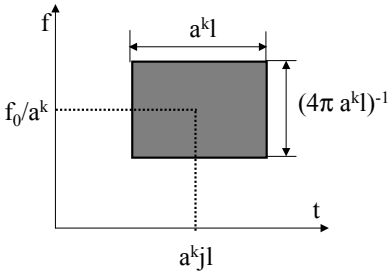
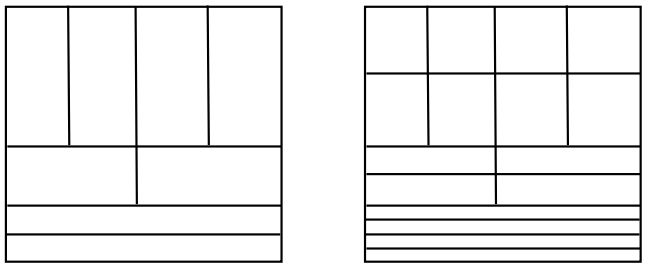


Fig. 3.24: Localisation et encombrement temps-fréquence d'une cellule élémentaire pour la transformée en ondelettes

Base d'ondelettes de Daubechies Le principe de la transformée en ondelettes consiste à représenter chaque fonction sur une base dont chaque élément est à support compact dans le temps ; la transformée de Fourier de chaque élément de la base est lui-même à support compact. Chaque élément de la base $\{\psi_{u,s}\}$ sera obtenu par dilatation d'une ondelette $\psi(t)$ par un facteur



a) b)

Fig. 3.25: Plan temps-fréquence partitionné en base d'ondelette : a) avec une voie par octave, b) avec deux voies par octave

d'échelle (équation 3.25). L'expression de la transformée en ondelettes est donnée par :

$$Wf(u,s) = \int_{-\infty}^{+\infty} f(t)\ \frac{1}{\sqrt{s}}\ \psi^*\left(\frac{t-u}{s}\right)\ dt. \tag{3.26}$$

Le facteur d'échelle comporte un aspect translation dans le temps grâce au terme u, et aussi par les termes s et $\sqrt{s}$ un aspect dilatation à la fois en temps et en amplitude. La dilatation en amplitude permet de conserver une norme constante pour tous les éléments de la base [Truchetet, 1998].

Le critère le plus important pour le choix d'une ondelette est de présenter, pour elle et sa transformée de Fourier, des oscillations les plus faibles possibles ; c'est ce qui assure une bonne résolution temporelle et fréquentielle.

Les ondelettes que l'on utilise souvent dans le cadre du traitement du signal mono-dimensionnel discret sont les ondelettes de Daubechies [Valens, 1999]. Pour la transformée rapide en ondelettes (DWT), les fonctions sont définies par un jeu d'indices que l'on désigne sous l'appellation «coefficients des filtres en ondelettes».

Les ondelettes de Daubechies à support compact sont décrites dans [Mallat, 2000]. Ce sont des fonctions à p moments nuls, leur régularité augmente avec p. Le nombre de coefficients est de 4 pour $p = 2$, de 12 pour $p = 6$ et de 20 pour $p = 10$. La forme de chaque ondelette pour 4, 12 et 20 coefficients est visualisée sur les Figures 3.26, 3.27 et 3.28.

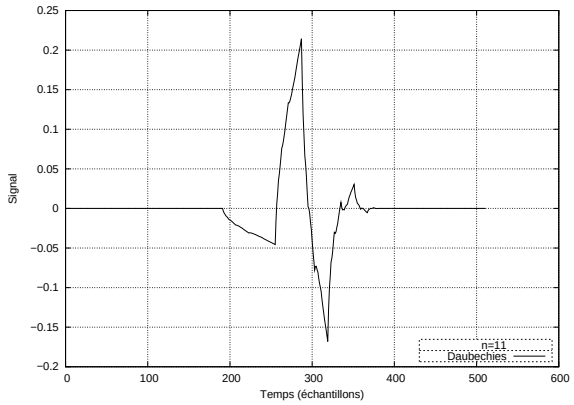


Fig. 3.26: Ondelettes de Daubechies à 2 moments nuls (4 coefficients).

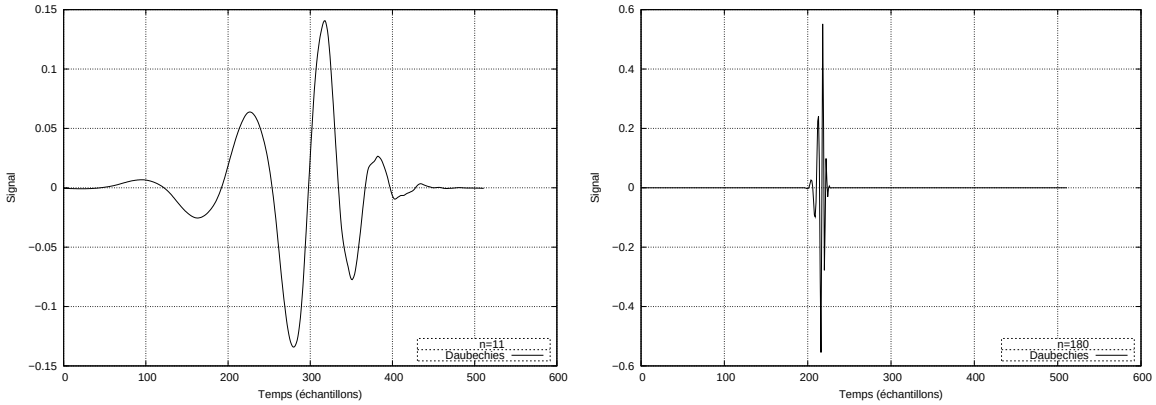


Fig. 3.27: Ondelettes de Daubechies à 6 moments nuls (12 coefficients) pour 2 facteurs d'échelle différents.

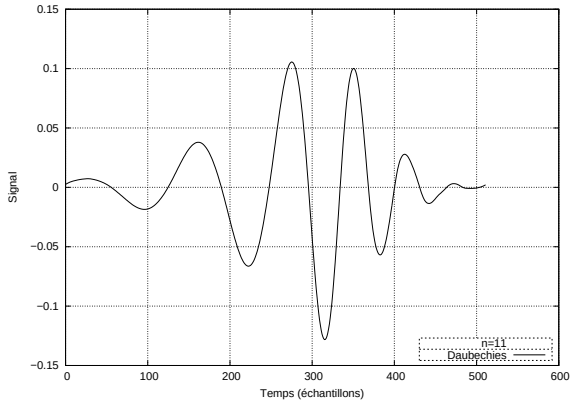


Fig. 3.28: Ondelettes de Daubechies à 10 moments nuls (20 coefficients).

Les coefficients d'ondelettes de DWT Nous utiliserons toujours des ondelettes de Daubechies dans la suite de l'étude pour le calcul de la transformée en ondelettes rapides DWT.

La transformée en ondelettes rapide se calcule, comme la TFR, sur une fenêtre de calcul contenant un nombre d'échantillons qui est une puissance entière de 2. Dans le cas de la TFR, la résolution étant constante, chaque élément calculé de la TFR $\hat{f}[n]$ correspond à la composante fréquentielle $n\frac{1}{2^N \delta t}$. En ce qui concerne la transformée en ondelettes, le pavage de l'espace temps-fréquence n'étant plus uniforme, le facteur d'échelle intervient [Mallat and Hwang, 1991].

Le résultat du calcul de la DWT d'une fenêtre de $2^N = 2^9$ échantillons, est un tableau de même taille. La Figure 3.29 montre la répartition des différents coefficients d'ondelettes en fonction des indices dans le tableau. Dans le cas d'une fenêtre de $1024 = 2^{10}$ échantillons, il y a ajout d'un onzième coefficient d'ondelettes indicé de 512 à 1024 et ainsi de suite. Le dernier coefficient d'ondelettes correspond à la moitié supérieure du tableau : sa taille est la moitié de la fenêtre d'analyse et il possède la résolution temporelle la plus fine ($2 \cdot \delta t$).

3.3.3.2 Algorithme de détection

L'algorithme se base sur la décomposition en ondelettes du signal à traiter. L'algorithme proposé commence avec le calcul de la transformée en ondelettes, suivi de celui de l'énergie de

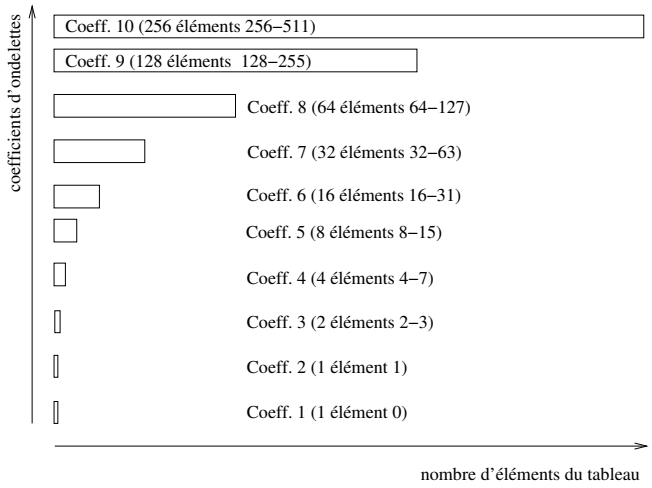


Fig. 3.29: Répartition des coefficients de la transformée en ondelettes dans le vecteur résultat.

chaque coefficient. Un exemple est donné à la Figure 3.31, pour un son de type claquement de porte superposé au bruit HIS, avec un RSB de 0 dB, représenté au haut de la figure. L'énergie des coefficients d'index 5 à 10 est également représentée. Le calcul de la transformée en ondelettes est effectué sur une fenêtre d'analyse de 512 échantillons, en utilisant les ondelettes de Daubechie de moment 6.

La part du bruit HIS dans l'énergie de chacun des coefficients est très forte pour le coefficient 5, elle décroît significativement à partir du coefficient 6 et devient presque négligeable à partir du coefficient 7. Le calcul de l'énergie des coefficients 7, 8, 9 ou 10 doit donc permettre de s'affranchir en grande partie du bruit HIS pour la détection. Une somme des énergies des coefficients 8 à 10 doit faire ressortir le son à détecter.

L'énergie des coefficients de la transformée en ondelettes pour le cas du bruit blanc a une valeur de fond (en l'absence de signal), celle-ci variant peu dans le temps pour les ordres de coefficients élevés (≥ 7). Les mêmes coefficients 8, 9 et 10 permettent la détection du son dans le cas du bruit d'écoulement de l'eau.

En conclusion, l'utilisation de l'énergie des coefficients élevés (de 8 à 10) de la transformée en ondelettes filtre bien les trois types de bruits et permet la détection des sons. L'utilisation d'un seuil adaptatif s'impose dans le cas du bruit blanc et du bruit d'écoulement d'eau.

Tenant compte des caractéristiques de l'énergie des coefficients de la transformée en ondelettes pour les types de bruit envisagés nous proposons l'algorithme avec l'organigramme présenté dans la Figure 3.30.

Les étapes de l'algorithme sont :

- Calcul de la transformée en ondelettes de la fenêtre d'analyse (2048 échantillons) $x_{DWT}(i)$
- Calcul de l'énergie de la somme des coefficients de 10 à 12 avec :

$$E = \frac{1}{512 - 64} \sum_{i=64}^{511} x_{DWT}^2(i)$$

- Calcul de la moyenne des L valeurs de l'énergie

- Seuillage de l'énergie avec seuil adaptatif λ

$$d(E) = \begin{cases} 1 \text{ si} & E \geq \lambda \\ 0 \text{ si} & E < \lambda \end{cases} \tag{3.27}$$

- Mise à jour du seuil adaptatif en fonction de la nouvelle valeur moyenne de l'énergie ($\lambda = \kappa + 1.2 * \mu$)

L'énergie des coefficients est calculée par une somme des carrés de tous les échantillons des coefficients concernés (de 10 à 12) et division par le nombre d'échantillons. C'est la valeur moyenne μ de l'énergie calculée sur 10 valeurs qui adapte le seuil. Le seuil adaptatif ne tient pas compte de l'écart-type mais il y a une marge de 20% de la moyenne pour tenir compte des variations rapides.

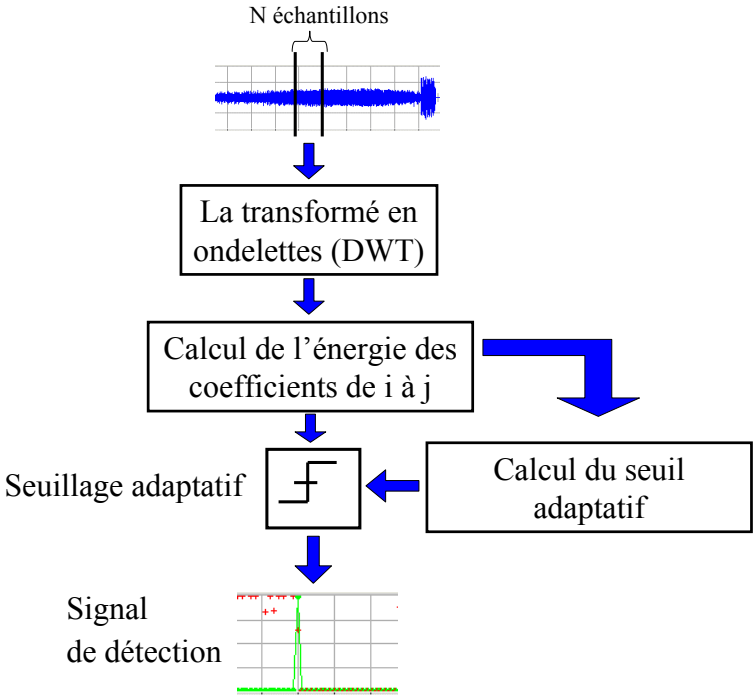


Fig. 3.30: Organigramme de l'algorithme basé sur les ondelettes

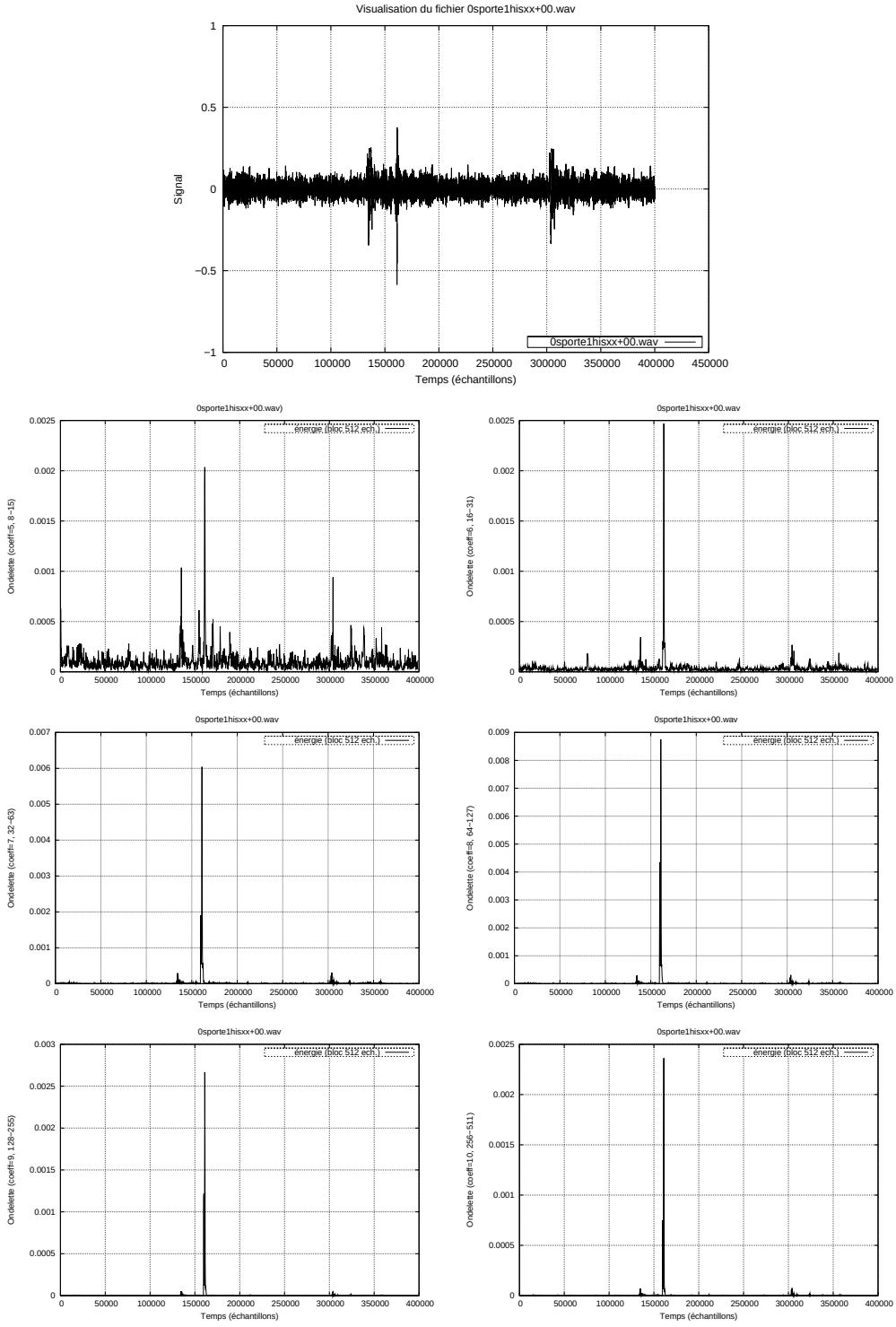


Fig. 3.31: Un claquement de porte superposé au bruit HIS avec un RSB de 0 dB. Variation temporelle de l'énergie des 5^e, 6^e, 7^e, 8^e, 9^e et 10^e coefficients d'ondelettes dans le cas du bruit HIS

3.4 Comparaisons

3.4.1 Corpus simulé (DSIM)

Les performances obtenues pour les trois algorithmes de détection proposés sont résumées dans le *tableau 3.3*. Dans le même tableau nous rappelons les performances des trois algorithmes issus de l'état de l'art évalué sur la même base de test. Les courbes ROC (taux de détections manquées en fonction des taux de fausses alarmes - **R**eceiver **O**perating **C**haracteristic) dans le cas du bruit environnemental HIS pour les trois algorithmes sont présentées dans l'annexe F.2 (page 170).

Méthode de détection	RSB [dB]	TEE pour les différents types de bruit de fond		
		Bruit blanc [%]	Bruit du HIS [%]	Bruit d'eau [%]
Algorithmes proposés				
Algo d'intercorrélation et normalisation par l'énergie de la fenêtre, recouvrement de 50%) (paragraph 3.3.1)	0	84.1	12.2	49.5
	+10	75.5	2.7	51.5
	+20	77.0	2.0	56.7
	+40	76.8	7.5	43.1
Algo de prédiction et seuillage adaptatif (paragraph 3.3.2)	0	30	62	30.8
	+10	7	28	15
	+20	7	7	15
	+40	0	0	15
Algo basé sur les ondelettes (paragraph 3.3.3)	0	**6**	**7.6**	**40**
	+10	**4**	**0**	**14.3**
	+20	**0**	**0**	**5**
	+40	**0**	**0**	**0**
Algorithmes issus de l'état de l'art				
Seuillage de la variance de l'énergie (paragraph 3.2.1.1)	0	5.7	69.6	41.9
	+10	**0**	**56.4**	**0.9**
	+20	0	33.4	0
	+40	0	5.2	6.1
Seuillage de l'énergie filtrée avec un filtre médian conditionné (paragraph 3.2.1.2)	0	30	64.9	24.1
	+10	**0**	**35.8**	**6.1**
	+20	0	10.4	0
	+40	0	0	0
Seuillage adaptatif de l'énergie (dépend d'un filtre médian) (paragraph 3.2.1.3)	0	11.4	64.2	31.5
	+10	**0**	**38.9**	**21.7**
	+20	0	31.8	1
	+40	0	15.7	0

Tab. 3.3: Résultats obtenus avec les trois algorithmes de détection proposés comparés à ceux des algorithmes issus de l'état de l'art

L'algorithme basé sur la fonction d'intercorrélation donne de très bons résultats avec le bruit HIS parce que ce bruit est fortement corrélé et l'apparition de l'événement sonore à détec-

ter réduit beaucoup l'intercorrélation entre les deux fenêtres successives d'analyse. Par contre le bruit blanc, par définition, est complètement décorrélé (deux segments consécutifs de signal sont indépendants et décorrélés pour un bruit blanc) alors l'apparition d'un événement sonore ne peut être détecté. En ce qui concerne le bruit d'eau qui a des caractéristiques de bruit coloré (bruit blanc dans une bande de fréquence) des performances moyennes sont obtenues. La remontée du TEE pour tous les types de bruit à RSB=40 dB s'explique par la normalisation du signal effectuée avant le calcul de la fonction d'intercorrélation. Dans le cas de bruits environnementaux fortement atténués ou absents, cette normalisation a tendance à fortement amplifier les très faibles signaux, qui ne sont pas nécessairement décorrélés. En conclusion ce type d'algorithme donne de très bons résultats sur le bruit de l'environnement d'étude.

L'algorithme basé sur la prédiction de l'énergie est utilisable pour des RSB égaux ou plus grands que 10 dB pour le bruit blanc et le bruit d'eau et de 20 dB pour le bruit HIS. Il a des bonnes performances sur l'eau et le bruit blanc parce que l'énergie de ces bruits varie lentement et l'algorithme peut la suivre (l'erreur de prédiction est petite) mais comme le bruit HIS a de fortes variations (surtout à des RSB de 0 et 10 dB) les performances sont très mauvaises (beaucoup de fausses détections).

L'algorithme fondé sur la décomposition en ondelettes du signal a de très bonnes performances pour le bruit HIS (5.6% de taux d'égale erreur pour un RSB de 0 dB et 0% pour toutes les autres valeurs de RSB) et pour le bruit blanc (7% pour un RSB de 0 et 0% pour tous les autres valeurs de RSB). Ces résultats s'expliquent par la bonne décomposition spectrale et temporelle obtenue par la transformée en ondelettes pour des sons impulsionnels. Un autre avantage, observable sur la courbe ROC de la Figure F.5, de cet algorithme, est le grand intervalle des seuils possibles pour obtenir ces performances (l'algorithme est moins sensible au seuil, respectivement aux gains du système d'acquisition). Dans le cas du bruit d'écoulement d'eau les performances sont inacceptables à RSB=0 dB et moyennes pour les autres valeurs de RSB. Cela s'explique par la composition spectrale du son d'écoulement d'eau par rapport à celle du bruit environnemental HIS.

Nous pouvons observer que les trois algorithmes issus de l'état de l'art ont de très bonnes performances en présence du bruit blanc mais des performances inutilisables dans la présence du bruit réel. L'algorithme basé sur la fonction d'intercorrélation est adapté seulement au bruit réel. L'algorithme de prédiction de l'énergie et celui basé sur la transformée en ondelettes ont de très bonnes performances avec le bruit réel et des performances comparables à celles des algorithmes issus de l'état de l'art pour le bruit blanc et celui de l'eau.

La Figure 3.32 présente les performances de quatre algorithmes selon Tanyer [Tanyer and Ozer, 2000] servant à la détection de la parole exposée à un bruit stationnaire. L'algorithme proposé par Tanyer a un TEE de 20% pour un RSB de 0 dB.

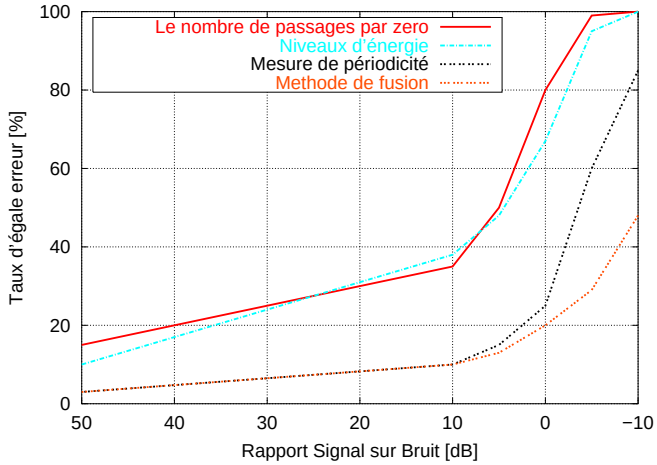


Fig. 3.32: Le Taux d'égale erreur en fonction du RSB pour quatre algorithmes de VAD d'après Tanyer 2000

3.4.2 Validation sur le corpus de sons pour la détection en conditions réelles (DREEL)

Pour confirmer les performances des algorithmes de détection obtenus sur le corpus de sons pour la détection qui a été créé avec des mélanges simulés, un corpus avec des enregistrements réels effectués dans l'appartement expérimental a été réalisé. Ce corpus contient 60 fichiers avec les mêmes sons que le corpus simulé, mais avec des RSB entre 2 dB et 30 dB (voir section 2.2.3 de la page 44). La moyenne du RSB est de 15 dB.

La Figure 3.33 présente les taux d'égale erreur des trois algorithmes proposés sur le corpus réel. Les valeurs des performances sur le corpus réel (DREEL) confirment les performances des algorithmes obtenus sur le corpus simulé (DSIM).

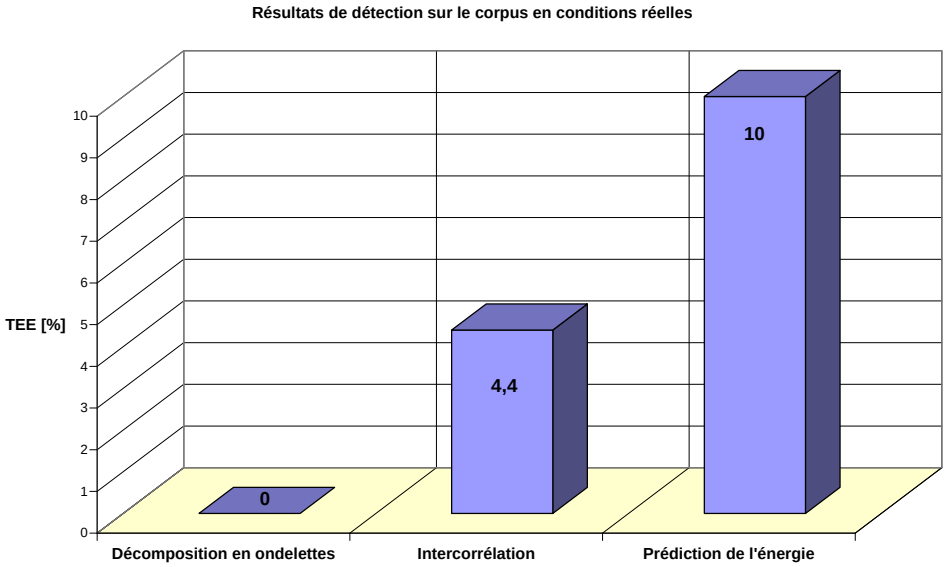


Fig. 3.33: Validation des résultats obtenus sur le corpus de sons en conditions réelles

3.5 Choix de l'algorithme tenant compte des contraintes induites dans un espace perceptif

Dans la section 3.2.1, trois algorithmes issus de l'état de l'art ont été présentés avec leurs performances sur le corpus de sons pour la détection. Ils ont de bonnes performances sur le bruit blanc mais sont inacceptables sur le bruit de l'environnement HIS. Dans la section 3.3, trois nouveaux algorithmes de détection sont proposés, avec des meilleurs performances dans les conditions réelles (bruit HIS). Pour le choix de l'algorithme, un compromis entre les performances et la rapidité de l'algorithme est fait.

Une estimation du temps de calcul de chaque algorithme sur un Pentium III 733MHz (avec une mémoire cache de 256 Ko) a été effectuée. Les résultats de cette estimation sont présentés dans la Figure 3.34. Les temps de calcul sont donnés en nombre de milliseconde nécessaires au traitement d'une seconde de signal. Il peut être intéressant pour fixer les idées d'exprimer la puissance de calcul nécessaire à chacun des algorithmes sur une échelle MIPS. Cependant, ces valeurs ne peuvent être qu'indicatives car une partie des calculs est faite en virgule flottante. Les algorithmes fondés sur la variance de l'énergie, le filtre médian et le seuil adaptatif nécessitent environ 1 MIPS, celui fondé sur la transformée en ondelettes 5.5 MIPS et celui fondé sur la fonction d'intercorrélation 30 MIPS. Ces résultats dépendent fortement de l'architecture du processeur.

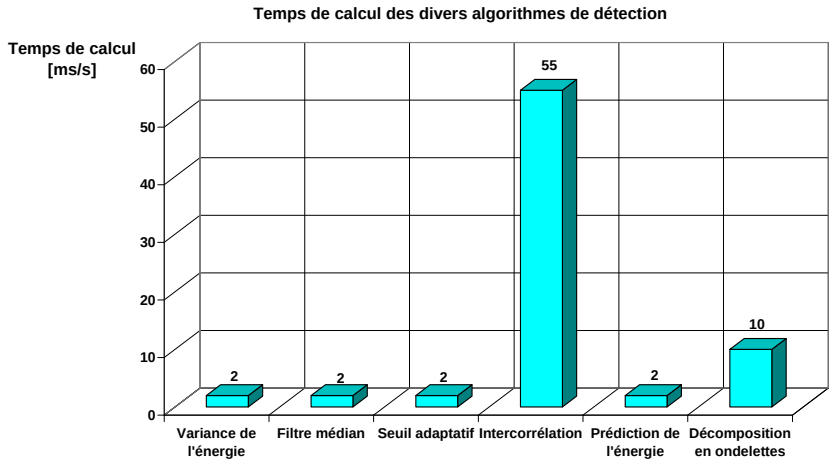


Fig. 3.34: Temps de calcul en ms/s des algorithmes de détection pour 1 s de signal à traiter

Le calcul de deux transformées de Fourier rapides (TFR) pour chaque fenêtre d'analyse du signal dans l'algorithme fondé sur la fonction d'intercorrélation explique le grand temps de calcul de cet algorithme. L'algorithme fondé sur la décomposition en ondelettes est 5 fois plus rapide que celui basé sur l'intercorrélation mais 5 fois moins rapide que tous les autres algorithmes qui sont très rapides en terme de temps de calcul.

Le Taux d'égale erreur (TEE) est une bonne indication des performances de l'algorithme mais pour une utilisation d'un algorithme de détection nous devons fixer une valeur du seuil. Le problème rencontré pour la majorité des algorithmes est que les valeurs du seuil pour obtenir le TEE sont différents d'un RSB à un autre. Comme en réalité la valeur du RSB est inconnue, la

Algorithme	Valeur du seuil fixe	RSB [dB]	TDM [%]	TFA [%]
Décomposition en ondelettes	0.005	0	**6**	**0**
		10	**0**	**0**
		20	**0**	**0**
		40	**0**	**0**
Intercorrélation	70	0	11	13
		10	4	32
		20	1	54
		40	2	57
Prédiction de l'énergie	4.5	0	25.3	72.4
		10	25.3	19.5
		20	6.1	0
		40	0	0

Tab. 3.4: Performances des trois algorithmes proposés pour un seuil fixe dans le cadre du bruit HIS

valeur du seuil doit être fixe. Cela explique les différences entre les performances présentées en terme de TEE des algorithmes et celles présentées dans le nouveau tableau 3.4. Parmi les trois algorithmes proposés, c'est l'algorithme fondé sur la décomposition en ondelettes, qui pour une valeur de seuil fixe conserve les mêmes performances. L'algorithme fondé sur l'intercorrélation conserve lui aussi des bonnes performances pour le TDM mais avec un TFA très grand. L'algorithme fondé sur la prédiction de l'énergie est inutilisable pour des valeurs de RSB en deçà de 20 dB.

En conclusion, l'algorithme fondé sur la décomposition en ondelettes a de très bonnes performances (TDM de 0% pour RSB $\geq$ 10 dB et de 6% pour RSB=0 dB) et un temps de calcul qui permet l'implémentation en temps réel. Tenant compte de ces performances nous avons décidé de l'implémenter dans l'environnement HIS.

3.6 Conclusions

Ce chapitre a présenté la détection des événements sonores dans le bruit. Après un état de l'art du domaine, trois algorithmes de détection issus de la littérature ont été présentés avec leurs performances. Ces algorithmes ont de bonnes performances sur du bruit blanc mais ils sont inutilisables dans le cas du bruit HIS (bruit de l'appartement expérimental) et du bruit d'écoulement d'eau.

En conséquence, trois nouveaux algorithmes de détection mieux adaptés au contexte d'un environnement perceptif sont proposés. L'algorithme de prédiction de l'énergie est très rapide mais il est utilisable seulement pour RSB $\geq$ 20 dB. L'algorithme fondé sur la fonction d'intercorrélation a de très bonnes performances pour le bruit HIS mais il nécessite un grand temps de calcul. L'algorithme fondé sur la décomposition en ondelettes a les meilleurs performances (TDM de 0% pour RSB $\geq$ 10 dB et de 6% pour RSB=0 dB) et un temps de calcul acceptable.

En conclusion, l'algorithme fondé sur la décomposition en ondelettes est la solution recherchée pour le système de détection et a été choisi pour être implémenté dans l'appartement d'étude.

Bibliographie

[Bellanger, 2002] Bellanger, M. (2002). *Traitement Numérique du Signal. Théorie et Pratique*. ISBN 2-10-006311-1. Dunod, Paris, 7ème edition.

[Boor, 1978] Boor, C. D. (1978). *A Practical Guides to SPLINES*. Springer-Verlag, New York.

[Bouvet, 1992] Bouvet, M. (1992). *Traitement des signaux pour les systèmes sonar*. ISBN 2-225-82615-3. Masson, Paris.

[Chang and Wu, 2000] Chang, D. C. and Wu, W. R. (2000). Feedback median filter for robust preprocessing of glint noise. *IEEE Transactions on Aerospace and Electronic Systems*, 22(6) :213–221.

[Colin, 2002] Colin, J. M. (2002). *Le Radar. Théorie et pratique*. ISBN 2-7298-1176-1. Ellipses,Paris.

[Dufaux, 2001] Dufaux, A. (2001). *Detection and recognition of Impulsive Sounds Signals*. PhD thesis, Faculté des sciences de l'Université de Neuchâtel, Suisse.

[Dufaux et al., 2000] Dufaux, A., Besacier, L., Ansorge, M., and Pellandini, F. (2000). Automatic sound detection and recognition for noisy environment. *European Signal Processing Conference (EUSIPCO), Tampere, Finlande*, pages 1033–1036.

[Duvaut, 1991] Duvaut, P. (1991). *Traitement du signal*. ISBN 2-86601-422-7. Hermès, Paris, 2ème edition.

[Gargour and Samir, 2001] Gargour, C. S. and Samir, C. (2001). *Traitement numérique des signaux*. ISBN 2-921145-24-3. Ecole de technologie supérieur, Canada.

[Gökhun and Özer, 2000] Gökhun, S. and Özer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 8(4) :478–482.

[Irwin, 1980] Irwin, M. J. (1980). Periodicity estimation in the presence of noise. *Acoustics Conference '80*, pages 243–247, Windemere, U.K.

[Junqua et al., 1991] Junqua, J. C., Reaves, B., and Mak, B. (1991). A study of endpoint detection algorithms in adverse conditions : Incidence on a DTW and HMM recognize. *Eurospeech '91*, pages 1371–1374, Genova, Italie.

[Kirsteins et al., 1997] Kirsteins, I. P., Mehta, S. K., and Fay, J. (1997). Power-law processors for detecting unkown signals in colored noise. *International Conference on Acoustics, Sounds and Signal Processing*, page 4pages, Munich, Allemagne.

[Kunt et al., 1991] Kunt, M., Bellanger, M., et al. (1991). *Techniques modernes de traitement numérique des signaux*, volume 1-3 of *ISBN 2-88074-207-2*. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Mallat, 2000] Mallat, S. (2000). *Une exploration des signaux en ondelettes*. ISBN 2-7302-0733-3. Les Editions de l'Ecole Polytechnique, Paris.

[Mallat and Hwang, 1991] Mallat, S. and Hwang, W. L. (1991). Singularity detection and processing with wavelets. Technical report, Courant Institute of Mathematical Sciences, New York University.

[Max, 1989] Max, J. (1989). *Traitement du Signal et Applications aux Mesures Physiques*, volume 1-4 of *ISBN 2-225-80470-2*. Masson, Paris, 4ème edition.

[Nemer et al., 2001] Nemer, E., Goubran, R., and Mahmoud, S. (2001). Robust voice activity detection using higher-order statistics in the lpc residual domain. *IEEE Transactions on Speech and Audio Processing*, 9(3) :217–231.

[Ozer and Tanyer, 1998] Ozer, H. and Tanyer, S. G. (1998). A geometric algorithm for voice activity detection in nonstationary gaussian noise. *EUropean SIgnal Processing COnference '98*, pages 543–547, Rhodes, Grèce.

[Press et al., 2002] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (2002). *Numerical Recipes in C ; The Art of scientific Computing ;The second Edition*. ISBN 0-521-43108-5. Cambridge University Press.

[Rabiner and Sambru, 1975] Rabiner, L. R. and Sambru, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *Bell Systems Technical Journal*, 54(2) :297–315.

[Seck, 2001] Seck, M. (2001). *Détection de ruptures et suivi de classe de sons pour l'indexation sonore*. PhD thesis, Université de Rennes I.

[Sohn et al., 1999] Sohn, J., Kim, N., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1) :1–3.

[Standard SAM, 1992] Standard SAM (1992). http ://www.icp.grenet.fr/relator/standsam.html.

[Tanyer and Ozer, 2000] Tanyer, S. G. and Ozer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 8(4) :478–482.

[Tepedelenlioglu et al., 2001] Tepedelenlioglu, C., Sidiropoulos, N., and Giannakis, G. (2001). Median filtering for power estimation in mobile communication systems. *IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications, Taoyouan, Taiwan*, pages 229–231.

[Truchetet, 1998] Truchetet, F. (1998). *Ondelettes pour le signal numérique*. ISBN 2-86601-672-6. Hermes.

[Valens, 1999] Valens, C. (1999). *A Really Friendly Guide to Wavelets*. http ://perso.wanadoo.fr/polyvalens/clemens/wavelets/wavelets.html.

[Wells et al., 1992] Wells, D., Barry, J., Grice, W., Fourcin, M., and Gibbon, A. (1992). SAM ESPRIT PROJECT 2589 - multilingual speech input/output assessment, methodology and standardisation. Final report. Technical Report SAM-UCL-G004, University CollegeLondon.

[Yamada and Watanabe, 2001] Yamada, T. and Watanabe, N. (2001). Voice activity detection using non-speech models and HMM composition. *Workshop on Hands-free Speech Communication, Tokyo, Japan*, pages 323–327.

[Zhang et al., 2002] Zhang, J., Ward, W., and Pellom, B. (2002). Phone based activity detection using online bayesian adaptation with conjugate normal distributions. *International Conference on Acoustics, Sounds and Signal Processing*, pages 123–127, Orlando, Floride, USA.

Chapitre 4

Classification des sons de la vie courante

4.1 Introduction à la classification des signaux sonores

Ce chapitre étudie la classification des sons de la vie courante. Les classes de sons à identifier, décrites dans le chapitre 2, ont été choisies en tenant compte des contraintes imposées par l'application de télésurveillance médicale.

La position de la classification dans le système d'analyse sonore est présentée en gras dans la Figure 4.1. Le rôle de la classification est de trouver l'appartenance d'un son, issu de l'algorithme de détection, à l'une des classes prédéfinies. Après l'identification du son, en fonction du type de la classe identifiée, le système décide s'il faut activer une alarme ou non. Comme le système de classification ne peut pas identifier le son avec une confiance de 100%, le son est envoyé aussi au centre de médico-surveillance, où il sera identifié par un auditeur humain.

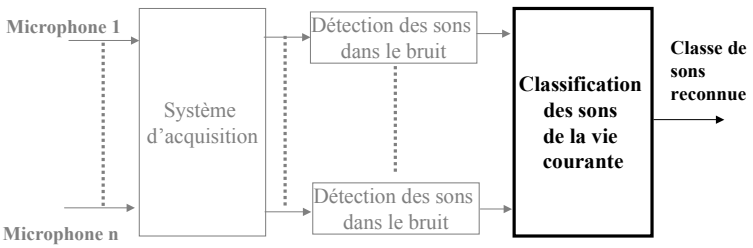


Fig. 4.1: Structure du système d'extraction d'informations sonores

Les techniques utilisées dans la reconnaissance des sons ne lui sont pas exclusives, car elles s'appliquent, sous le nom générique de «*Reconnaissance des Formes*», à nombre de signaux et d'images [Kunt et al., 2000]. Nous pouvons schématiser la «*Reconnaissance des Formes*» en la présentant comme la discipline de la «perception automatique».

Les techniques de reconnaissance de formes peuvent être divisées en *statistiques* et *structurelles*. Les méthodes *statistiques* modélisent la classe des formes à reconnaître par ses propriétés statistiques et identifient une nouvelle forme par calcul d'une probabilité d'appartenance. Dans l'approche *structurelles*, une classe est décrite par une grammaire. Une grammaire est composée d'un ensemble de règles syntaxiques qui déterminent l'ensemble des formes admissibles

dans cette classe et présentant en principe des caractéristiques structurelles communes.

Pour la classification des formes en général et pour celle des sons en particulier, deux cas se présentent en fonction de la connaissance *a priori* des sons à classifier. Nous parlons d'*apprentissage non supervisé* quand le système effectue une classification aveugle des sons (le corpus utilisé pour l'apprentissage n'est pas étiqueté). Dans ce cas il est possible d'utiliser l'algorithme des K-moyennes des voisins les plus proches (K-means Nearest Neighbors) [Theodoridis and Koutroumbas, 1998]. Lorsque le corpus d'apprentissage est étiqueté nous parlons d'*apprentissage supervisé*. C'est en l'occurrence notre cas.

Jusqu'à présent, dans le domaine de la classification des formes, des efforts ont été faits pour le développement de systèmes capables de reconnaître des signaux provenant de sources spéciales comme la parole ou les signaux sonar ou radar, mais moins d'efforts pour des systèmes capables de détecter, isoler et classifier des sons de la vie courante [Seck et al., 2001]. En effet, dans l'espace des sons qui est défini comme étant l'espace des ondes sonores se propageant dans l'air et ayant une bande de fréquences limitée à [20 Hz - 20 kHz], la plupart des recherches ont été dédiées à la reconnaissance de la parole. Les autres sons, à l'exception de la musique, ont reçu moins d'attention car ils étaient souvent ignorés ou éliminés.

Le terme «*classification*» désigne l'identification de quelques aspects particuliers du son ou de sa source. Par exemple, pour la *musique* la classification peut être l'identification d'un instrument musical dans une partition et/ou la transcription des notes qu'il joue. Pour la *parole*, la classification désigne la reconnaissance des mots prononcés pour une application de reconnaissance de la parole ou l'identification du locuteur pour une application de reconnaissance du locuteur.

Une application de reconnaissance des sons est constituée de deux modules principaux : une partie d'extraction des paramètres acoustiques les plus significatifs à partir du signal et une partie de reconnaissance de formes qui prend la décision d'identification du son.

Les applications de reconnaissance des sons peuvent être classées en fonction des paramètres acoustiques utilisés et en fonction du type d'algorithme de classification. Parmi les paramètres acoustiques on distingue : des bancs de filtres, MFCC[1]/LFCC[2], LPC[3]/LPCC[4], des paramètres issus des statistiques d'ordre supérieur, des paramètres issus de la transformée de Fourier à court terme, etc. Parfois l'extraction des paramètres acoustiques est suivie d'une réduction de données de type quantification vectorielle (VQ), analyse en composantes principales (PCA) ou analyse discriminante linéaire (LDA). Les méthodes les plus souvent utilisées de classification incluent : les méthodes statistiques (HMM, GMM, etc.), les réseaux neuronaux et les systèmes experts à base de règles.

Les applications les plus importantes pour la classification des signaux sonores sont les suivantes, notre application se situant dans la catégorie «classification des sons environnementaux» :

- Reconnaissance automatique de la parole
- Traitement du signal pour les sonars
- Classification de sons biologiques et biomédicaux

[1] Mel-Frequencies Cepstral Coefficients
[2] Linear Frequencies Cepstral Coefficients
[3] Linear Prediction Coefficients
[4] Linear Prediction Cepstral Coefficients

- Transcription musicale automatique
- Analyse de scènes (identification des sources sonores dans un flux sonore)
- Indexation automatique de bandes sonores
- **Classification de sons environnementaux**

La reconnaissance automatique de la parole a comme but la transcription textuelle du signal de parole. La réalisation d'un système de reconnaissance de parole est nécessaire pour l'interface vocal homme-machine. Ces interfaces ont beaucoup d'applications comme : système de dictée pour un traitement de texte, système de commande vocale de divers appareils ou dans les télécommunications. En fonction de la complexité de la tâche nous avons des systèmes de reconnaissances des mots isolés ou de la parole continue, avec vocabulaire restreint ou large ou seulement pour l'identification de mots clés (Word Spotting).

Sonar (**SO**und **NA**vigation and **R**anging) désigne une méthode ou un équipement pour déterminer la présence, la localisation et la nature des objets dans l'eau par l'utilisation des sons. Il existe deux types de sonars : le sonar actif qui est similaire au radar (il transmet un signal acoustique qui est réfléchi par la cible) et le sonar passif (il identifie l'objet par l'analyse des sons propres aux objets). Ces systèmes sont utilisés par les sous-marins pour l'identification du relief sous-marin ou l'identification de l'ennemi dans le cas des applications militaires.

La classification des sons biologiques et biomédicaux est utilisée pour la surveillance et/ou le diagnostic des patients. De tels systèmes ont été développés pour la reconnaissance de signaux issus du poumon ou du coeur.

La transcription musicale automatique décompose la musique en notes et a comme applications possibles la génération des partitions musicales d'un orchestre à partir d'un enregistrement ou le pilotage d'un synthétiseur par un instrument.

L'analyse de scènes est la reconnaissance de diverses sources sonores d'un enregistrement. Ces systèmes s'inspirent de l'audition et de l'analyse humaine et leur but est la transcription d'un signal sonore en une liste de descriptions symboliques des sons (la reconnaissance de la parole et la transcription musicale en sont un cas particulier) et la séparation du signal en plusieurs en fonction de la source.

L'indexation automatique de bandes sonores est la segmentation d'un document sonore en sous-parties. Par exemple, pour un signal de parole l'indexation peut être faite en fonction du locuteur ou du contenu linguistique. Pour la musique une indexation en fonction du type de musique, de l'instrument ou même du chanteur, peut être faite.

La classification de sons environnementaux a comme but l'identification de quelques classes de sons de l'environnement. Parmi les applications éventuelles nous avons : la classification des voitures en fonction de leur bruit, l'identification des sons des armes à feu pour avertir la police, l'identification des sons de chute pour les systèmes de télésurveillance médicale.

Dans un premier temps, notre méthodologie a consisté à adapter les techniques utilisées dans le domaine de la reconnaissance de la parole et/ou du locuteur à la classification des sons de la vie courante. Nous avons choisi d'utiliser des techniques de classification statistiques pour la classification des sons de la vie courante parce que les caractéristiques de ces sons peuvent être modélisées facilement par un processus statistique.

Ce chapitre continue avec un état de l'art des techniques de classification statistiques et des paramètres acoustiques. Des premiers résultats sont présentés avec des paramètres acoustiques classiquement utilisés en reconnaissance de la parole et une méthode de classification à base de

GMM (Gaussian Mixture Models). La section suivante présente de nouveaux paramètres acoustiques plus spécifiques parmi lesquels des paramètres issus de la transformée en ondelettes. Une étude statistique de la pertinence des paramètres acoustiques en rapport avec le corpus de sons de la vie courante est aussi présentée. La section se termine par la présentation des résultats obtenus avec les nouveaux paramètres proposés.

La dernière section étudie l'influence de la présence du bruit sur le taux moyen de classification. Pour améliorer les performances de classification deux possibilités sont proposées : le recours à des paramètres robustes au bruit et le débruitage du signal avant la classification.

4.2 Analyse de l'existant dans le domaine de la classification des sons

Un système de classification statistique modélise chaque classe de signaux avec une répartition statistique (le plus souvent des gaussiennes). Les étapes de la classification statistique sont : la paramétrisation acoustique du signal suivie du calcul de la vraisemblance du signal en rapport avec les modèles des classes à identifier, conformément à la Figure 4.2 [CALLIOPE, 1989]. La sortie du système est la classe d'appartenance la plus probable.

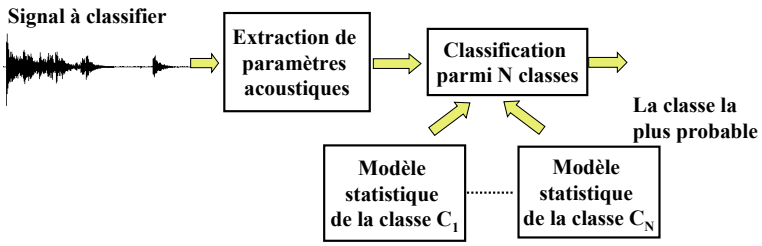


Fig. 4.2: Structure générale d'un système de classification sonore statistique

Le signal à classifier ne peut pas être utilisé directement parce que l'information qu'il contient est à l'état brut avec beaucoup de redondance. Dans la Figure 4.3 nous pouvons observer que le signal d'une sonnerie dans le domaine fréquentiel a une structure avec harmoniques qui le différencie des autres signaux [Woodard, 1992]. Par la transformation du signal temporel en une collection de paramètres acoustiques nous réduisons la quantité d'informations.

Parmi les techniques de classification les plus utilisées dans le domaine de la reconnaissance de la parole/du locuteur nous avons :

- Les modèles de Markov cachés (HMM - Hidden Markov Models)
- Les modèles de mélange de gaussiennes (GMM Gaussian Mixture Models)
- L'alignement temporel dynamique (DTW - Dynamic Time Warping)

Les modèles de Markov cachés résultent de l'association d'un ensemble de fonctions de densité de probabilité (ou distributions de probabilité) et d'une chaîne de Markov [Rabiner and Juang, 1993]. Les fonctions de densité de probabilité donnent les probabilités sur l'ensemble des observations acoustiques et la chaîne de Markov sert de support aux distributions. Un modèle de Markov peut être défini par un automate probabiliste d'états finis [Jouvet, 1988].

Les GMM modélisent la distribution des paramètres acoustiques par un mélange de gaussiennes. Cette méthode se fonde sur l'hypothèse qu'un mélange de gaussiennes peut représenter

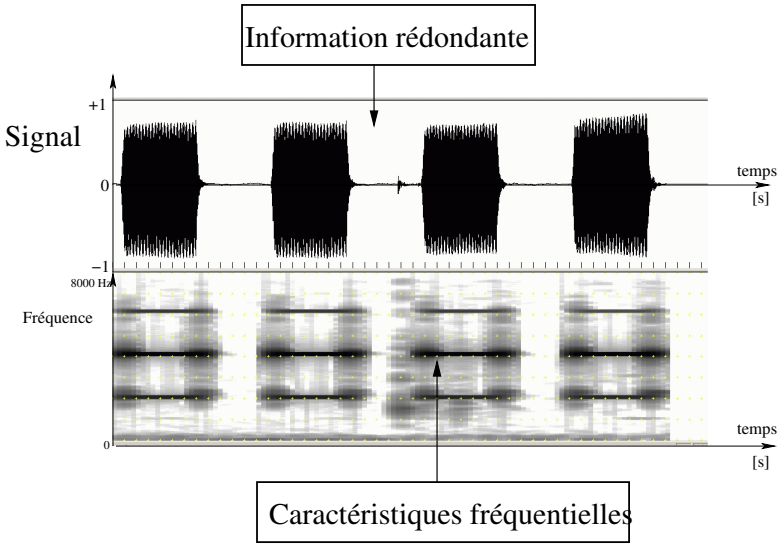


Fig. 4.3: Passage de la représentation temporelle d'un signal à la représentation temps-fréquence

n'importe quel ensemble de paramètres acoustiques [Reynolds, 1994]. Dans la Figure 4.4 nous avons un exemple de modélisation d'une fonction de répartition (en gras sur la figure) par 4 gaussiennes. Une application à la segmentation parole/musique est présentée dans la référence [El-Maleh et al., 2000].

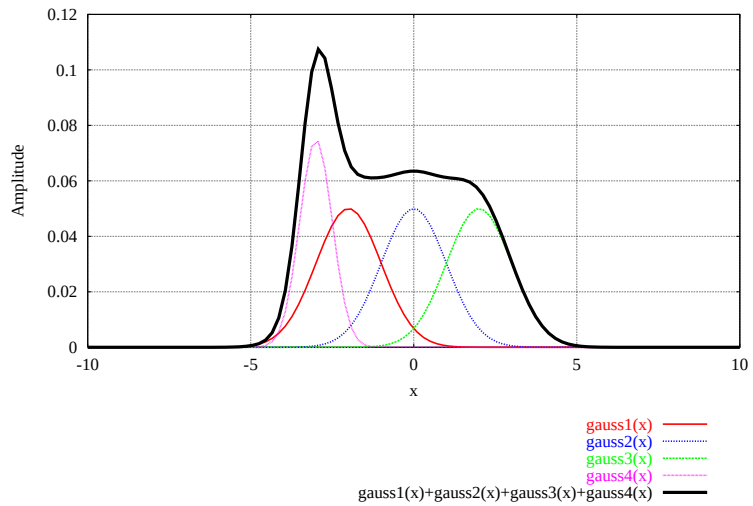


Fig. 4.4: Modèle de mélange de 4 gaussiennes : en noir le mélange des 4 gaussiennes

L'alignement temporel dynamique(DTW) trouve le chemin optimal (optimal dans le sens de la dissemblance minimale des chemins) d'un vecteur de paramètres sur une grille en deux dimensions de tous les vecteurs connus. L'axe horizontal de la grille DTW correspond aux vecteurs des paramètres d'apprentissage $X = (x_1, x_2, \ldots, x_N)$ et l'axe vertical au vecteur des paramètres à identifier $Y = (y_1, y_2, \ldots, y_T)$ comme illustré dans la figure 4.5. Les points «start» et «stop» représentent les points entre lesquels le chemin est cherché. Au final, le signal d'entrée est associé à l'état connu pour lequel la dissemblance du chemin optimal est minimale. Une

application à la reconnaissance de la parole de DTW est décrite en [Myers and Rabiner, 1981], et une amélioration de l'algorithme est proposée en [Yaniv and Burshtein, 2003].

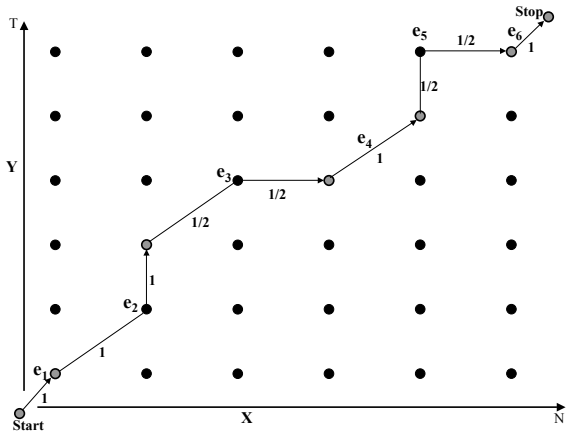


Fig. 4.5: La grille de DTW et le chemin valide imposé par les contraintes

Parmi le faible nombre d'études disponibles sur la classification des sons, Goldhor propose un système à base de GMM pour la classification de 23 sons [Goldhor, 1993]. Les sons considérés sont : son d'une alarme, aboiement d'un chien, son d'une balle (2 types), son d'une sonnette, son d'une bouteille en verre, son de ventilateur (2 types), son de moteur de voiture, sonnerie de porte, son d'une cloche électrique, son d'une perceuse manuelle, sonnerie mécanique de téléphone, sonnerie électrique de téléphone (2 types), claquement de porte (2 types), écoulement de l'eau en évier et en baignoire, son d'aspirateur, le son d'un violon, sifflet d'une bouilloire. La base d'apprentissage et de test contient 268 signaux. La paramétrisation est effectuée avec des coefficients cepstraux et un système de type GMM est utilisé pour la classification. Le meilleur taux de bonne reconnaissance est de 98%. Cependant, cette étude est limitée à un petit corpus dans des conditions non-bruitées.

Une autre méthode de classification des sons environnementaux est celle des produits de deux modèles HMM proposée par Woodard [Woodard, 1992]. Cette méthode utilise un HMM pour la modélisation de forme spectrale et autre pour le gain. La décision de classification est prise sur le produit des scores des deux modèles HMM. Trois classes de sons naturels sont considérées : ouverture et fermeture d'une porte en bois, un outil de métal qui chute dans un récipient métallique et écoulement de l'eau d'un récipient à un autre. Le taux de bonne reconnaissance obtenu est de 96%. Étude limitée à trois types de sons non-bruités.

Papadopoulos décrit un système temps réel pour la détection, la poursuite, l'estimation de la localisation en deux dimensions et la caractérisation spectrale des sources sonores à large bande de fréquences [Papadopoulos et al., 1992]. Le système utilise plusieurs antennes acoustiques pour la localisation ; un système est proposé pour la classification, basé sur la comparaison entre le spectre normalisé acquis et une liste de spectres associée à des sons connus. Seulement trois types de sons sont considérés : une série d'impulsions, le son d'une perceuse et le son d'un aspirateur. Des résultats quantitatifs ne sont pas donnés mais les auteurs considèrent qu'il est possible de caractériser les sons de façon suffisamment discriminante.

La classification des sons de sonnerie (téléphone, alarme d'horloge, sonnerie de porte, alarme de feu, etc.) est étudiée par Paradie [Paradie and Nawab, 1990]. Deux modèles de sons

de sonneries sont introduits : un modèle pour les signaux de sonneries sinusoïdales et un deuxième pour les signaux de sonneries qui peuvent être modélisés par le filtrage d'un bruit coloré.

Cowling présente une comparaison entre les méthodes ANN (réseaux neuronaux), DTW et VQ pour la classification de 8 types de sons environnementaux [Cowling and Sitte, 2002]. Les résultats d'un test sur un corpus de seulement 35 s sont donnés. La méthode VQ a un taux de bonne reconnaissance de 62%.

Un système de classification des sons d'hélicoptères, d'avions et des trains est proposé dans [Cabell and Fuller, 1989] et [Scott et al., 1993]. L'algorithme se base sur un arbre des classes appliqué à un jeu réduit de paramètres extraits de la transformée de Fourier et de la fonction d'autocorrélation. Un algorithme de réduction des paramètres (LDA) est utilisé pour réduire les 108 paramètres initiaux à 4 - 10 paramètres. La base totale (test et apprentissage) contient seulement 162 signaux. Le meilleur taux de bonne reconnaissance obtenu est de 90%.

Pour notre étude, nous avons choisi d'utiliser un système à base de modèles de mélange de gaussiennes (GMM) et de l'adapter à la classification des sons de la vie courante. Ce choix a été fait pour plusieurs raisons : la modélisation par un mélange de gaussiennes est très flexible par rapport au type de signal et les GMM représentent un bon compromis entre les performances et la complexité. Les HMM sont complexes, en demandant de plus longs temps de calcul et sont moins adaptés à la classification des signaux courts. Selon Dufaux ([Dufaux, 2001]) les performances obtenues avec les HMM qui divisent le son en trois parties, ne sont pas meilleures que celles obtenues avec un système fondé sur les GMM.

La section suivante présente le modèle de mélange de gaussiennes (GMM).

4.2.1 Le modèle de mélange de Gaussiennes GMM

La classification de sons à l'aide d'un modèle GMM comprend 2 étapes : une phase d'apprentissage du système sur un ensemble de fichiers supposés représentatifs d'une classe et, une deuxième phase, de vérification de l'appartenance d'un son quelconque à cette classe. La phase d'apprentissage d'une classe est initiée uniquement si une modification a été apportée dans la définition de cette classe : ajout ou suppression de fichiers sons caractéristiques de la classe. Pendant la phase d'apprentissage, la modélisation statistique des paramètres acoustiques du son est effectuée. La répartition des paramètres acoustiques d'une classe de son dans l'espace est modélisée par une somme de fonctions de densités de probabilités gaussiennes.

A chaque instant d'échantillonnage, par exemple toutes les 16 ms, le système évalue les d paramètres acoustiques correspondant au signal audio qui vient d'être acquis. L'ensemble de ces d paramètres constitue le vecteur acoustique.

Pour estimer le rapport de vraisemblance d'un vecteur acoustique, nous avons utilisé une base de distributions multi-gaussiennes, c'est-à-dire que la distribution d'observations appartenant à une même classe de sons est modélisée par une somme pondérée de M distributions gaussiennes [Boite et al., 2000]. Cela revient à considérer que les vecteurs observés, x_i, sont des réalisations de variables aléatoires mutuellement indépendantes dont la densité de probabilité $f_m(x_i)$ est de type gaussienne. La formule 4.1 donne la modélisation de la distribution d'observations $f(x_i)$ par une somme pondérée par les coefficients π_m des distributions gaussiennes

$f_m(x_i)$.

$$f(x_i) = \sum_{m=1}^{M} \pi_m \cdot f_m(x_i)$$

$$\text{où :} \tag{4.1}$$

$$\pi_m \geq 0, \forall m \in [1, M] \qquad \text{et} \qquad \sum_{m=1}^{M} \pi_m = 1$$

La densité de probabilité gaussienne a la forme de l'équation (4.2).

$$f_m(x_i) = \frac{e^{\left[-\frac{1}{2}(\overrightarrow{x}-\overrightarrow{\mu_m})^t C_m^{-1}(\overrightarrow{x}-\overrightarrow{\mu_m})\right]}}{(2\pi)^{\frac{d}{2}}\sqrt{\det(C_m)}} \tag{4.2}$$

où :

- $\overrightarrow{x}$ est le vecteur des distributions à modéliser
- $\overrightarrow{\mu_m}$ est le vecteur moyen du vecteur $\overrightarrow{x}$
- C_m la matrice de covariance du vecteur $\overrightarrow{x}$
- d est la dimension du vecteur $\overrightarrow{x}$
- C^t est la matrice C transposé et C^{-1} est la matrice C inversé

Lors de la phase d'apprentissage, tous les vecteurs acoustiques d'une même classe de sons sont utilisés pour déterminer le poids correspondant à chacune des N gaussiennes, le vecteur acoustique moyen et la matrice de covariance de chacune des gaussiennes. Le vecteur acoustique moyen et la matrice de covariance se réduisent respectivement à la moyenne et à l'écart-type dans le cas d'une distribution gaussienne mono-dimensionnelle.

Pour chacune des gaussiennes ($1 \leq m \leq M$) de la classe ω_k, les paramètres suivants sont ceux qui caractérisent le modèle GMM de la classe :

- le nombre de paramètres acoustiques utilisés d (toujours le même)
- les poids de chaque gaussienne $\pi_{k,m}$ qui respecte la condition : $\sum_{m=1}^{M} \pi_{k,m} = 1$
- les vecteurs moyens $\mu_{k,m}$
- les matrices de covariance $C_{k,m}$

4.2.1.1 Apprentissage

L'apprentissage a pour but d'estimer les paramètres des gaussiennes qui composent le modèle à partir des vecteurs acoustiques des sons compris dans la classe. L'apprentissage d'une classe se décompose en deux étapes successives : tout d'abord l'obtention de valeurs approximatives des paramètres des gaussiennes de la classe par l'algorithme des K-moyennes (ou «K-means»), ensuite l'optimisation des valeurs de ces paramètres par un algorithme de type EM (**E**xpectation **M**aximisation).

Algorithme des K-moyennes L'algorithme des K-moyennes cherche à regrouper l'ensemble des vecteurs x_j d'apprentissage d'une classe en N sous-ensembles disjoints [Rabiner and Juang, 1993]. Cet ensemble de N sous-ensembles sera appelé dictionnaire. Chaque

sous-ensemble est caractérisé par son barycentre ou centroïde. Cet algorithme n'est que localement optimal, il est donc influencé par ses conditions initiales. La variante LBG (voir [Linde et al., 1980]) de cet algorithme comporte 4 étapes et elle sera décrite ensuite. L'optimisation qui est cherchée par l'algorithme, consiste en la réduction de la distance euclidienne entre chaque élément d'un sous-ensemble et le centroïde du sous-ensemble. Le nombre de sous-ensembles est la taille K du dictionnaire et ce sera une puissance entière de 2, $K = 2^p$. Les étapes de cet algorithme sont les suivantes :

1. **Initialisation.** Le dictionnaire est constitué d'un seul sous-ensemble contenant tous les vecteurs d'apprentissage, son centroïde est la moyenne des vecteurs d'apprentissage. A cette étape transitoire, $K = 1$ (ce n'est pas encore une puissance de 2).
2. **Éclatement du dictionnaire.** Chaque sous-ensemble du dictionnaire va être éclaté en remplaçant chaque centroïde de coordonnées y_i par 2 nouveaux centroïdes de coordonnées respectives $y_i(1+\varepsilon)$ et $y_i(1-\varepsilon)$, avec $\varepsilon \ll 1$ (valeur possible : $\varepsilon = 0.01$). La valeur de K est doublée par rapport à la valeur précédente.
3. **Optimisation du dictionnaire.** Chaque vecteur x_j sera examiné à tour de rôle. Dans une première étape la distance euclidienne séparant x_j et chacun des centroïdes est calculée ; le vecteur x_j étant alors affecté au sous-ensemble pour lequel cette distance est la plus faible. Lorsque chaque vecteur a été affecté à un sous-ensemble, la moyenne des vecteurs de chaque sous-ensemble est recalculée pour obtenir le centroïde correspondant.
   Cette étape doit être réitérée plusieurs fois avant de passer à l'étape 4.
4. **Test d'arrêt.** Tant que $K < 2^p$, le dictionnaire est à nouveau éclaté et optimisé en réitérant les étapes 2 et 3. Sinon, le dictionnaire a atteint la taille désirée et pourra alors être optimisé au moyen de l'algorithme EM.

Algorithme EM L'algorithme EM fait intervenir des variables latentes que l'on ne peut observer directement. Dans notre cas, chaque vecteur x_j est décrit non seulement par les d paramètres acoustiques (valeurs mesurées) mais aussi par le sous-ensemble S_i (défini par un centroïde) auquel il se rattache. Nous avons noté que dans le cas de l'algorithme LBG, l'hypothèse avait été faite que chaque x se rattachait réellement à un sous-ensemble. Dans le cas de l'algorithme EM ce ne sera plus le cas. Celui-ci va maximiser la vraisemblance de façon itérative, mais le vecteur x sera maintenant rattaché aux M sous-ensembles S_i avec une probabilité particulière, sans que l'on puisse déterminer à quel sous-ensemble S_i il appartient réellement. C'est ce paramètre que l'on qualifie de donnée *cachée* ou *latente*, voir [Boite et al., 2000] et [Cappé, 2001].

L'idée de base de l'algorithme EM consiste à raisonner sur les données observées et latentes, tout en prenant en compte le fait que l'information disponible sur les données latentes provient des données observées. A chaque étape k de l'algorithme, le calcul de la variable latente, pour chaque x_i et chaque gaussienne $\Theta_{k,m}$ des coefficients est effectué conformément à l'équation (4.3) où $|C_m|$ est le déterminant de la matrice C_m.

$$\begin{aligned} \gamma_i^{(k)}(m) &= P(S_i = m \mid x_i; \Theta_{k,m}) \\ &= \frac{\pi_m \cdot |C_m|^{-1/2} \cdot e^{\left[-\frac{1}{2}(x_i-\mu_m)^t \cdot C_m^{-1} \cdot (x_i-\mu_m)\right]}}{\sum_{j=1}^{M} \pi_j \cdot |C_j|^{-1/2} \cdot e^{\left[-\frac{1}{2}(x_i-\mu_j)^t \cdot C_j^{-1} \cdot (x_i-\mu_j)\right]}} \end{aligned} \tag{4.3}$$

Ce calcule utilise les paramètres $\Theta_{k,m}$ des gaussiennes déterminées à l'étape précédente $(k-1)$ et permet le calcul de la quantité intermédiaire $\mathcal{Q}_{\Theta_k}(\Theta)$ que l'on doit maximiser avec l'équation (4.4).

$$\mathcal{Q}_{\Theta_k}(\Theta) = \sum_{t=1}^{T}\sum_{i=1}^{M} log\big(\pi_i f_i(x_t)\big) \cdot \gamma_t^{(k)}(i) \tag{4.4}$$

La ré-estimation des paramètres $\Theta_{k+1,m} = (\pi_m^{(k+1)}, \mu_m^{(k+1)}, \Sigma_m^{(k+1)})$ à partir des paramètres $\Theta_{k,m}$ constitue la deuxième étape de l'algorithme EM. La maximisation de (4.4) par rapport aux paramètres π_m, μ_m et C_m fournit les nouvelles valeurs estimées des paramètres [Cappé, 2001] pour l'itération $(k+1)$. Les formules de calcul des paramètres à l'itération $(k+1)$ sont données en (4.5).

$$\begin{cases} \pi_m^{(k+1)} &= \frac{1}{T}\sum_{t=1}^{T}\gamma_t^{(k)}(m) \\ \mu_m^{(k+1)} &= \dfrac{\sum_{t=1}^{T}\gamma_t^{(k)}(m)\cdot x_t}{\sum_{t=1}^{T}\gamma_t^{(k)}(m)} \\ C_m^{(k+1)} &= \dfrac{\sum_{t=1}^{T}\gamma_t^{(k)}(m)\cdot(x_t-\mu_m^{(k+1)})\cdot(x_t-\mu_m^{(k+1)})^{\mathcal{T}}}{\sum_{t=1}^{T}\gamma_t^{(k)}(m)} \end{cases} \tag{4.5}$$

4.2.1.2 Classification

Pendant la phase de classification, on doit déterminer la classe ω_l la plus probable à partir du calcul de la vraisemblance [Dufaux, 2001], [Boite et al., 2000], pour le vecteur acoustique x obtenu à l'instant t, et pour chacune des classes de sons ω_k $(1 \leq k \leq K)$:

$$p(x \mid \omega_k) = \sum_{m=1}^{M} \pi_{k,m} \cdot \frac{1}{(2\pi)^{\frac{d}{2}}|C_{k,m}|^{\frac{1}{2}}} \cdot e^{\left[-\frac{1}{2}(x-\mu_{k,m})^{\mathcal{T}}\cdot C_{k,m}^{-1}\cdot(x-\mu_{k,m})\right]} \tag{4.6}$$

En pratique, il est nécessaire de déterminer la vraisemblance d'un son constitué d'une suite temporelle de n vecteurs x_i, $(x_i)_{1\leq i\leq n}$. Elle peut être obtenue à partir de la vraisemblance de chacun des vecteurs x_i selon l'équation (4.7).

$$p(X \mid \omega_k) = \prod_{i=1}^{n} p(x_i \mid \omega_k) \tag{4.7}$$

Un signal à tester est transformé dans une suite de n vecteurs acoustiques X qui ont d paramètres acoustiques. Il appartiendra avec le maximum de vraisemblance à la classe ω_l pour laquelle $p(X \mid \omega_l)$ est maximale, conformément à l'équation (4.8).

$$p(X \mid \omega_l) = \max_{k=1}^{K} \left(p(X \mid \omega_k) \right) \tag{4.8}$$

4.2.2 La paramétrisation du signal

L'apprentissage est effectué non pas directement sur les signaux temporels mais sur des paramètres extraits de ceux-ci parce que le signal temporel contient beaucoup d'informations redondantes. Le passage à une représentation fréquentielle du signal met en évidence les caractéristiques du signal. Par exemple dans la Figure 4.3 (de la page 99), la représentation fréquentielle d'un signal de sonnerie téléphonique montre la présence de trois harmoniques qui par leurs caractéristiques décrivent le signal de sonnerie [Pfeiffer et al., 1996].

Les étapes de calcul des paramètres acoustiques, représentées dans la Figure 4.6 sont : le fenêtrage du signal avec une fonction spécifique par une fenêtre glissante et le calcul proprement dit des paramètres acoustiques. Parmi les différentes fonctions de fenêtrage, les plus utilisées sont : la fenêtre rectangulaire, la fenêtre de Hamming, la fenêtre de Hanning et la fenêtre de Blackmann. En traitement de la parole, la fenêtre de Hamming est la plus utilisée pour sa bonne résolution de séparation des harmoniques d'amplitudes rapprochées.

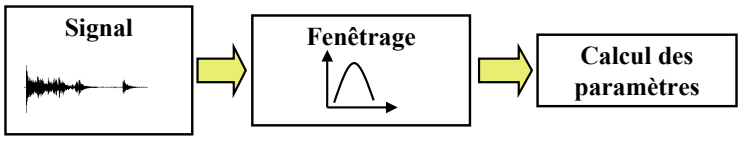


Fig. 4.6: Extraction des paramètres

Les fenêtres d'analyse ont généralement une longueur temporelle de 20 ms et se superposent de 50% en vue de lisser la variation des paramètres acoustiques.

Après le fenêtrage du signal suit le calcul des paramètres acoustiques [Boite et al., 2000]. En reconnaissance de la parole les paramètres utilisés classiquement sont les MFCC, △MFCC, △△MFCC, LPC. Pour les sons de la vie courante, il n'y a pas encore eu d'étude pour trouver les paramètres acoustiques les mieux adaptés. C'est pour cette raison que nous avons décidé de concentrer notre étude sur une recherche des paramètres acoustiques les plus efficaces. Les paramètres classiques, comme MFCC, LPC sont testés tout d'abord. De nouveaux paramètres sont ensuite proposés.

4.2.2.1 Les paramètres acoustiques classiques

L'énergie du signal Comme paramètre acoustique, on utilise l'énergie logarithmique du signal qui est définie comme suit :

$$E = \ln \left(\sum_{i=0}^{N} s_i^2 \right) \tag{4.9}$$

où N est le nombre d'échantillons du signal, s_i l'échantillon i du signal qui a une valeur comprise dans l'intervalle $[-32768, 32767]$.

Coefficients d'énergie Trois types de coefficients d'énergie ont été étudiés :

- **Rectangle** : La moyenne de l'énergie spectrale sur une bande de fréquences obtenue par un filtrage rectangulaire ;
- **Triangle** : La moyenne de l'énergie spectrale sur une bande de fréquences obtenue par un filtrage triangulaire ;
- **Mel** : La moyenne de l'énergie spectrale sur une bande de fréquences Mel.

Pour les coefficients en bandes rectangulaires et en bandes triangulaires, un filtrage du spectre de puissance du signal est fait avec des filtres rectangulaires (respectivement triangulaires) et la valeur moyenne de la puissance sur cette bande constitue le paramètre acoustique. En ce qui concerne les coefficients en fréquence Mel, des filtres triangulaires centrés sur des fréquences Mel sont utilisés.

LPC (Linear Prediction Coefficients) & LPCC (Linear Prediction Cepstral Coefficients) Les coefficients LPC sont basés sur le modèle de production de la parole, qui considère que l'appareil de production de la parole (cordes vocales et conduit vocal complet) est constitué d'une source (source pseudo-périodique ou source de bruit) et d'un filtre se comportant comme un résonateur (conduit vocal). La Figure 4.7 schématise ce modèle simplifié de la production de parole. Le signal de parole peut être ainsi modélisé comme étant le signal en sortie d'un filtre H(z) dont la source d'excitation à l'entrée du filtre u(n) est soit une source de série d'impulsions quasi-périodiques, soit une source de bruit aléatoire.

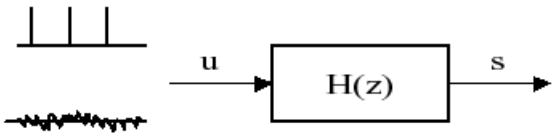


Fig. 4.7: Modèle source-filtre de production de la parole

L'analyse LPC repose sur l'hypothèse que le filtre est un filtre tous-pôles (la formule 4.10).

$$H(z) = \frac{S(z)}{G.U(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \tag{4.10}$$

où G est le coefficient de gain, a_k sont les coefficients LPC et p est l'ordre du filtre.

Avec cette hypothèse, le signal de la parole peut être considéré comme un signal autorégressif :

$$s(n) = \sum_{k=1}^{p} a_k.s(n-k) + G.u(n) \tag{4.11}$$

Les coefficients a_k et le gain G sont calculés grâce à des méthodes fondées sur le calcul de la matrice de covariance ou grâce à des méthodes fondées sur le calcul de la matrice d'auto

corrélation (la méthode utilisée est basée sur la matrice d'auto corrélation décrite en annexe A.1).

Les coefficients LPCC (c_n) sont dérivés directement des coefficients LPC à travers le système d'équations suivant :

$$\begin{cases} c_0 = \ln G \\ c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} & 1 \leq m \leq p \\ c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} & m > p \end{cases} \tag{4.12}$$

où G est le coefficient de gain du modèle source-filtre.

MFCC (Mel Frequency Cepstral Coefficients) Les coefficients MFCC sont des coefficients cepstraux très souvent utilisés en reconnaissance automatique de la parole. Le calcul des paramètres MFCC utilise une échelle fréquentielle non-linéaire qui tient compte des particularités de l'oreille humaine [Davis and Mermelstein, 1980].

L'échelle de fréquence Mel (B) est définie par (4.13).

$$B(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{4.13}$$

où f représente la fréquence en Hz et $B(f)$ la fréquence suivant l'échelle de fréquence Mel.

La procédure de calcul des coefficients MFCC est présentée dans la Figure 4.8.

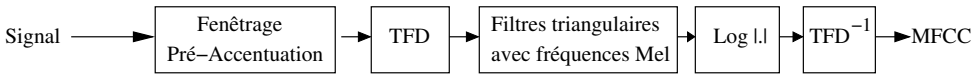


Fig. 4.8: Calcul des MFCC

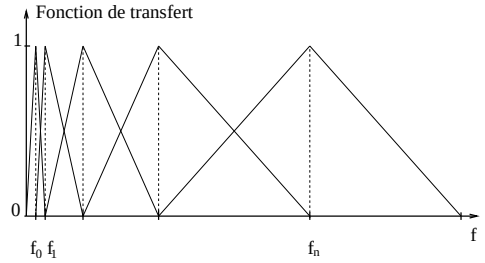
Soit un signal discret $s(n)$ avec $0 \leq n \leq N-1$, N est le nombre d'échantillons d'une fenêtre analysée, F_s est la fréquence d'échantillonnage, la transformé de Fourier discrète $S(k)$ est obtenue avec la formule (4.14).

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{\frac{-j2\pi nk}{N}} \quad \text{avec} \quad 0 \leq k < N \tag{4.14}$$

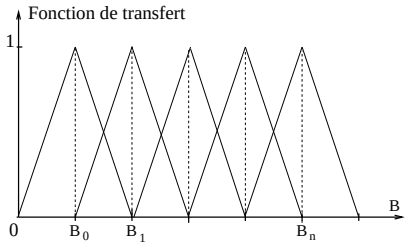
Le spectre du signal est filtré par des filtres triangulaires (voir Figure 4.9) dont les bandes passantes sont de même largeur dans le domaine des fréquences Mel. Les points de frontières B_m des filtres en échelle de fréquence Mel sont calculés à partir de la formule (4.15).

$$B_m = B_1 + m\frac{B_h - B_b}{M+1} \quad 0 \leq m \leq M+1 \tag{4.15}$$

où M désigne le nombre de filtres, f_h la fréquence la plus haute et f_b la fréquence la plus basse du signal.



(a) Les filtres triangulaires passe-bande en fréquence de base f



(b) Les filtres triangulaires passe-bande en fréquence Mel, B_f

Fig. 4.9: Filtres en fréquences Mel et uniforme

Dans le domaine fréquentiel, les points f_m discrets correspondants sont calculés d'après (4.16).

$$f_m = \left(\frac{N}{F_s}\right) B^{-1}\left(B_b + m\frac{B_h - B_b}{M+1}\right) \tag{4.16}$$

où $B^{-1}(i)$ désigne la fréquence correspondante à la fréquence i de l'échelle Mel, $B_i^{-1} = 700(10^{\frac{i}{2595}} - 1)$

Les coefficients cepstraux de fréquence en échelle Mel (MFCC) peuvent être obtenus par une transformée de Fourier inverse à partir des coefficients en sorties des filtres. Mais le nombre de MFCC est moins grand que le nombre de filtres, donc on utilise plutôt une transformée en cosinus discrète (équation 4.17).

$$c(n) = \begin{cases} \dfrac{1}{\sqrt{M}} \displaystyle\sum_{m=0}^{M-1} E(m) & , n = 0 \\ \sqrt{\dfrac{2}{M}} \displaystyle\sum_{m=0}^{M-1} E(m) \cos\left(\dfrac{\pi n(m+\frac{1}{2})}{M}\right) & , 0 \leq n < M \end{cases} \tag{4.17}$$

LFCC (Linear Frequency Cepstral Coefficients) Les coefficients LFCC sont calculés de la même manière que les MFCC, mais avec la différence que les fréquences des filtres sont uniformément réparties sur l'échelle linéaire des fréquences, et non plus sur une échelle Mel.

Les dérivées des coefficients (Δ, $\Delta\Delta$) Pour pouvoir tenir compte des variations dans le temps des paramètres pour les GMM qui calculent une valeur de vraisemblance à chaque instant on utilise les dérivées de ceux-ci. La dérivée d'un paramètre acoustique est la mesure de sa variation dans le temps. Comme la fonction de variation des paramètres acoustiques est inconnue et que l'on n'en connaît que des valeurs à des instants précis, le calcul de la première dérivée et de la deuxième dérivée se fait par une approximation. Les formules d'approximation du calcul de la première et de la deuxième dérivée sont données dans l'annexe A.2 (page 152).

4.2.3 Étude statistique des paramètres acoustiques classiques

Le choix des paramètres acoustiques est déterminant pour la bonne attribution d'une classe à un signal. Pour pouvoir analyser et choisir les paramètres les mieux adaptés pour la reconnaissance des sons, ont été calculés pour chaque paramètre et chaque classe des sons : la valeur

moyenne, la variance, l'écart-type, les valeurs maximales et minimales. Les histogrammes de répartition des valeurs d'un paramètre pour chaque classe sont déterminés aussi.

Enfin, pour choisir les paramètres les mieux adaptés à notre corpus de classes de sons (CPUR), une synthèse de ces résultats a été faite en calculant le rapport de Fisher FDR (Fisher Discriminant Ratio) qui est décrit dans la section suivante.

La meilleure combinaison des paramètres acoustiques non corrélés pourrait être établie en testant toutes ces combinaisons et en évaluant les taux de bonne classification obtenus, mais cette méthode est très coûteuse en temps et donc peu utilisée. Sinon, des protocoles comme : «La sélection séquentielle en avant » (Sequential Forward Selection - SFS) [Couvreur, 1997], l'algorithme de Viterbi [Viterbi, 1967] ou l'algorithme sous-optimal «Add-on» [Goldstein, 1976] peuvent être utilisés. Tous ces algorithmes consistent en l'addition d'un paramètre acoustique après un autre jusqu'au M^{e} paramètre, en conservant à chaque étape le nouveau paramètre qui maximise le critère de performances. Il y a aussi, le protocole inverse, «la sélection séquentielle en arrière» (Sequential Backward Selection - SBS) qui démarre avec tous les paramètres et enlève le plus mauvais, un par un.

Une autre possibilité pour la sélection des paramètres acoustiques parmi ceux calculés pour chaque trame est d'appliquer une transformation linéaire sur chaque vecteur acoustique. Cette transformation peut décorréler les paramètres et rehausser leurs capacités discriminantes. Les principales méthodes sont : l'analyse en composantes principales (PCA) et l'analyse discriminante linéaire (LDA). La transformation du vecteur acoustique est effectuée comme suit à l'aide d'une matrice A :

$$\overrightarrow{p_{tr}} = A.\overrightarrow{p} \tag{4.18}$$

où $\overrightarrow{p_{tr}}$ est le vecteur acoustique transformé et $\overrightarrow{p}$ est le vecteur acoustique initial.

Pour l'analyse en composantes principales la matrice A prend la forme donnée en 4.19 [Jolliffe, 1986].

$$A = [\overrightarrow{u_1}, \overrightarrow{u_2}, \ldots, \overrightarrow{u_D}]^t \tag{4.19}$$

où $\overrightarrow{u_i}$ contient les vecteurs propres de la matrice de covariance ordonnés de manière croissante (sa dimension est D_r - la dimension de l'espace réduit des paramètres) et D est la dimension du vecteur acoustique.

Pour l'analyse discriminante linéaire la matrice A est appelée «la matrice de covariance de Fisher". Cette matrice est égale au rapport entre la matrice de dispersion interclasses S_b et la matrice de dispersion intraclasse S_w, conformément à l'équation (4.20). La matrice est calculée pour maximiser ce rapport.

$$A = S_w^{-1}.S_b \tag{4.20}$$

Toutes ces méthodes de sélection des paramètres nécessitent des grands temps de calcul, mais elles permettent une optimisation multidimensionnelle. Autrement, on peut obtenir une évaluation du caractère discriminant de chaque paramètre acoustique à l'aide du critère connu comme «Le rapport de Fisher» (Fisher Discriminant Ratio) ou avec «La mesure multimodale de chevauchement» (Multimodal Overlap Measure - MOM) [Kil and Shin, 1996].

4.2.3.1 La mesure multimodale de chevauchement (MOM)

La pertinence des paramètres est déterminée par ce critère en intégrant la surface de chevauchement entre les distributions des fonctions de probabilité des classes sur tout l'espace des

paramètres. Les paramètres avec un degré de chevauchement réduit donne une grande valeur pour ce critère [Häcker et al., 2002]. Le MOM pour chaque paramètre X_p est défini comme suit :

$$\text{MOM} = \int J(p(X_p|C_1), \ldots, p(X_p|C_k))dX_p \tag{4.21}$$

où :

- J est une fonction qui mesure le degré de chevauchement du paramètre dans l'espace des fonctions de distribution de probabilité $p(X_p|C_i)$
- $p(X_p|C_i)$ est la fonction de distribution de probabilité du paramètre X_p pour la classe C_i

La fonction J peut avoir plusieurs formes. La plus simple est d'utiliser le minimum des fonctions de distribution de probabilités :

$$J(p) = 1 - \sum_{i=1}^{k} \sum_{j \neq i}^{k} \int \min(p(X_p|C_i), p(X_p|C_j))dX_p \tag{4.22}$$

Une autre approche pour la fonction J est le produit de toutes les fonctions de distribution de probabilité pour le même paramètre :

$$J(p) = 1 - \prod_{i=1}^{k} p(X_p|C_i) \tag{4.23}$$

4.2.3.2 Le critère de Fisher (FDR)

Ce critère estime la capacité de chaque paramètre à distinguer différentes classes, en mesurant le chevauchement de leurs fonctions de densité de probabilité.

Le critère de Fisher, pour des fonctions de densité de probabilité gaussiennes, peut être calculé pour chaque paramètre comme suit :

$$\text{FDR} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} (\overline{x[i]} - \overline{x[j]})^2}{\sum_{i=1}^{k} Var(x)[i]} \tag{4.24}$$

où $\overline{x[i]}$ désigne la moyenne du paramètre x pour la classe i et $Var(x)[i]$ la variance du paramètre x pour la classe i.

Pour chaque dimension du paramètre, il représente le rapport entre la distance qui sépare deux classes i, j et leurs variances. Dans la formule (4.24), les contributions de toutes les K classes sont cumulées. Ce paramètre peut être interprété comme étant le rapport de la variabilité interclasse du paramètre par la variabilité intraclasse du même paramètre.

Avec ce critère, les paramètres avec les meilleurs potentiels de pertinence peuvent être sélectionnés. Le désavantage de ce critère est que, généralement, il n'intègre pas les relations de corrélation entre les paramètres. L'utilisation de ce critère donne un jeu de paramètres «sous-optimal». En [Kil and Shin, 1996], sur un corpus donné, la différence des performances de classification obtenues avec le critère de Fisher et celles obtenues par un critère d'optimisation multidimensionnelle est de 8%.

4.2.3.3 Résultats statistiques

A titre d'exemple, nous présentons, ci-dessous, en détail les résultats obtenus pour le *deuxième coefficient MFCC*, de même qu'une synthèse des résultats pour tous les autres paramètres classiques. Le tableau 4.1 indique la moyenne, l'écart-type, les valeurs minimale et maximale du deuxième coefficient MFCC pour chaque classe de signal. Ce paramètre arrive à bien séparer les classes de sons ayant des moyennes pour chaque classe bien différentes. Par exemple, la classe 7 - Vaisselle est bien séparée de la classe 9 - Serrure de porte par ce coefficient.

Classe	Deuxième coefficient MFCC			
	Moyenne	Écart-type	Minimum	Maximum
C1 - Claquement de porte	9.35	6.15	-20.21	31.32
C2 - Bris de verre	-0.60	5.03	-19.80	15.70
C3 - Sonneries de téléphone	-5.44	7.33	-27.96	10.13
C4 - Sons de pas	7.07	4.14	-13.63	25.32
C5 - Cris	3.63	10.15	-26.75	24.25
C7 - Vaisselle	2.65	6.69	-19.35	22.72
C9 - Serrures de porte	-13.94	7.06	-24.91	7.36

Tab. 4.1: Statistiques du deuxième coefficient MFCC pour toutes les classes

La Figure 4.10 présente la valeur moyenne avec l'écart-type du deuxième coefficient MFCC pour l'ensemble des classes. Cette figure permet de visualiser la capacité de ce paramètre à discriminer les classes de sons.

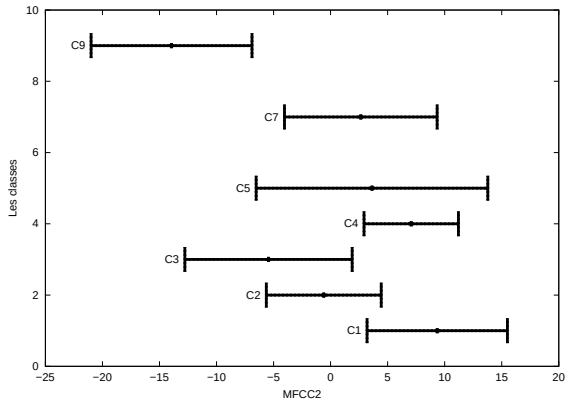


Fig. 4.10: Séparation des classes par le deuxième coefficient MFCC

La valeur du rapport de Fisher pour le deuxième coefficient MFCC est $FDR = 16.07$. Ce calcul confirme les déductions résultant de l'analyse de la moyenne et de l'écart-type de ce paramètre en rapport avec chaque classe de sons.

Le tableau 4.2 indique les valeurs des rapports de Fisher pour les paramètres acoustiques classiques.

Parmi les valeurs du tableau 4.2, si l'on considère seulement celles qui dépassent la valeur 7, l'on peut conclure que les coefficients les plus pertinents sont :

- **MFCC** : seulement les coefficients 2, 3 et 4 sont pertinents

Paramètre	FDR	Paramètre	FDR	Paramètre	FDR	Paramètre	FDR
MFCC1	2.72	LFCC1	5.34	**LPC1**	**20.42**	LPCC1	4.57
MFCC2	**16.07**	**LFCC2**	**18.55**	LPC2	4.50	**LPCC2**	**17.25**
MFCC3	**10.33**	**LFCC3**	**8.88**	LPC3	5.73	**LPCC3**	**20.33**
MFCC4	**10.02**	**LFCC4**	**7.50**	LPC4	3.35	LPCC4	5.98
MFCC5	2.01	**LFCC5**	**7.90**	LPC5	3.87	LPCC5	5.84
MFCC6	2.91	**LFCC6**	**7.91**	LPC6	1.65	**LPCC6**	**7.94**
MFCC7	3.36	LFCC7	3.91	LPC7	0.90	**LPCC7**	**11.36**
MFCC8	3.60	**LFCC8**	**7.77**	LPC8	1.16	**LPCC8**	**7.43**
MFCC9	0.53	LFCC9	4.53	LPC9	2.11	LPCC9	5.04
MFCC10	3.34	LFCC10	6.40	LPC10	3.15	LPCC10	6.16
MFCC11	2.88	LFCC11	4.22	LPC11	2.38	LPCC11	6.39
MFCC12	3.20	LFCC12	1.40	LPC12	2.45	LPCC12	4.34
MFCC13	1.48	LFCC13	2.32	LPC13	1.97	LPCC13	4.11
MFCC14	3.61	LFCC14	2.58	LPC14	1.97	LPCC14	4.87
MFCC15	3.26	LFCC15	3.53	LPC15	0.87	LPCC15	4.57
MFCC16	4.41	LFCC16	2.03	LPC16	0.33	LPCC16	4.28

Tab. 4.2: Valeurs FDR pour les paramètres MFCC, LFCC, LPC et LPCC

- **LFCC** : les coefficients 2-6 et 8 sont pertinents
- **LPC** : seulement le premier coefficient semble être pertinent
- **LPCC** : les coefficients 2, 3, 6, 7 et 8 sont pertinents

Les paramètres LPC ont été définis en tenant compte du modèle de production de la parole (dans le cas de la parole, le signal des cordes vocales est filtré par le conduit vocal et les coefficients LPC mesurent les pôles du filtre) ce qui justifie les mauvais résultats obtenus dans le cas des sons de la vie courante qui ne peuvent pas être modélisé par le même modèle de production.

4.2.4 Résultats de la classification avec paramètres acoustiques classiques

4.2.4.1 Les algorithmes et la méthodologie utilisés

La plate-forme ELISA [Chagnolleau et al., 2001] issue principalement des travaux du laboratoire LIA[5] d'Avignon est utilisée. Cette plate-forme nous a permis de tester des modèles de mélange de gaussiennes (GMM) avec des matrices de covariance diagonales.

La durée de la fenêtre de calcul des paramètres acoustiques a été choisie de 16 ms avec un recouvrement de 50%.

La première étape des calculs consiste à déterminer le modèle GMM, à partir de tous les fichiers d'une classe de son. Ensuite, chacun des fichiers à tester est évalué en fonction de chaque modèle GMM calculé précédemment (figure 4.11). La moyenne du taux de vraisemblance est calculée sur la durée totale du son à classifier. Le taux de vraisemblance est une probabilité, avec des valeurs comprises dans l'intervalle $[0, 1]$, de sorte que son domaine de variation exprimé en échelle logarithmique devient $(-\infty, 0]$. Ce son à classifier est finalement alloué à la classe pour laquelle la moyenne est la plus forte.

[5] Laboratoire Informatique d'Avignon

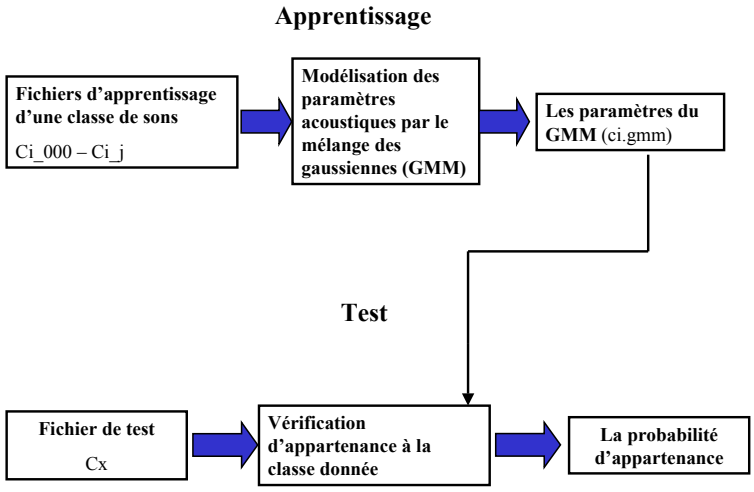


Fig. 4.11: Processus de reconnaissance

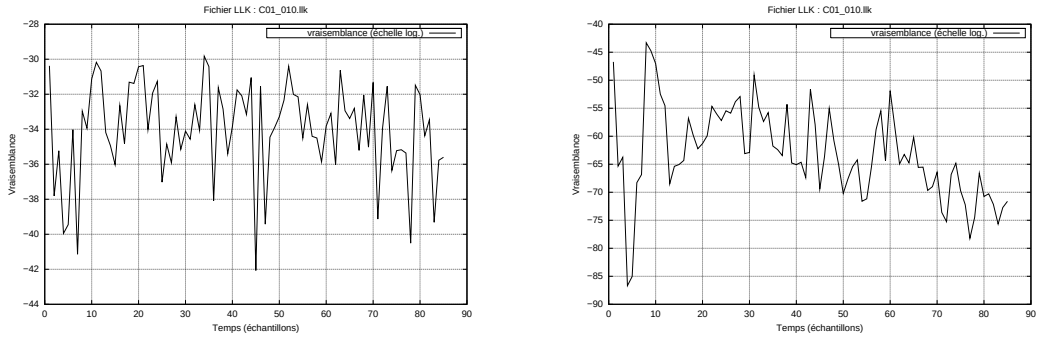
Le test d'appartenance d'un son à une classe donne une variation temporelle de la vraisemblance logarithmique. Le modèle GMM ne tient pas compte de la variation temporelle du signal, il modélise la répartition de l'ensemble des paramètres acoustiques d'une classe de sons. En conclusion, le système calcule la moyenne géométrique des valeurs de vraisemblance des trames du signal (voir équation (4.25)). La moyenne la plus grande indique la classe d'appartenance, voir équation (4.26).

$$\log_{10}(p(X \mid w_k)) = \log_{10}\left(\prod_{i=0}^{n} p(x_i \mid w_k)\right) \tag{4.25}$$

où $p(X \mid w_k)$ est la vraisemblance d'appartenance du son X à la classe w_k, n le nombre de trames du signal X et $p(x_i \mid w_k)$ la vraisemblance d'appartenance de la trame x_i du signal X à la classe w_k.

$$w_{l\ \text{reconnue}} = w_k \mid \log_{10}(p(X \mid w_k)) \text{ est maximal} \tag{4.26}$$

Le premier graphique de la Figure 4.12 présente la variation temporelle de la vraisemblance d'un son de la classe C1 rapportée au modèle de la classe C1. Les valeurs de vraisemblance varient entre -30 et - 42, avec une moyenne de -34. Le deuxième graphique de la même figure 4.12 présente la variation dans le temps de la vraisemblance du même son, rapportée au modèle de la classe C5. Les valeurs de vraisemblance varient entre -45 et - 85, avec une moyenne de -62. En conclusion, ce fichier est considéré comme appartenant plus probablement à la classe C1 qu'à la classe C5.



En rapport avec le modèle de la classe C1 En rapport avec le modèle de la classe C5

Fig. 4.12: Variation temporelle du logarithme de la vraisemblance du fichier C010 de la classe C1 avec le modèle respectivement le modèle des classes C1 et C5

4.2.4.2 Le Protocole de test

Les performances d'un algorithme de classification dépendent beaucoup des données utilisés en apprentissage et en test. Le corpus de données est limité et il doit être partitionné dans une partie d'apprentissage et une autre de test. Parmi les protocoles de test existants, nous pouvons citer [JAIN et al., 2000] :

- **Leave all in** utilise tous le corpus pour l'apprentissage et en même temps pour le test. Il assure un bon apprentissage mais il produit une vue optimiste des performances de l'algorithme. Ce protocole ne semble pas vraiment rigoureux pour une évaluation.
- **La validation croisée** utilise une partie du corpus pour l'apprentissage et l'autre pour le test. Parmi les protocoles les plus utilisés, on trouvera :
  - **Holdout Techniques** Les parties de test et d'apprentissage sont fixées au début. Cependent, ceci peut être problématique lorsqu'on dispose d'un corpus de petite taille.
  - **Leave one out** utilise tous les données sauf une pour l'apprentissage. Il permet d'utiliser un maximum de données pour l'apprentissage et il est très utilisé lorsque les corpus sont de taille insuffisante.
- **Ré-échantillonnage** divise le corpus en une partie d'apprentissage et de test aléatoirement. Cette procédure donne des corpus de taille arbitraire.

La procédure de test utilisée pour nos expériences est de type «leave-one-out» (spécifique au corpus de taille réduite) ce qui veut dire qu'à tout instant la base d'apprentissage comprend tous les fichiers sauf un, qui est le fichier de test. Les étapes de la procédure de test sont :

1. Un son est éliminé de l'ensemble d'apprentissage
2. Le calcul des modèles GMM de chaque classe est effectué
3. Le test du son éliminé au pas précédent est réalisé en rapport avec les modèles de toutes les classes. On trouve la classe d'appartenance la plus probable
4. Réintroduction du son dans l'ensemble d'apprentissage avec retrait d'un autre son. Si il n'y a pas d'autres sons à éliminer on s'arrête, sinon on revient au début.

Ce protocole est illustré par la figure 4.13.

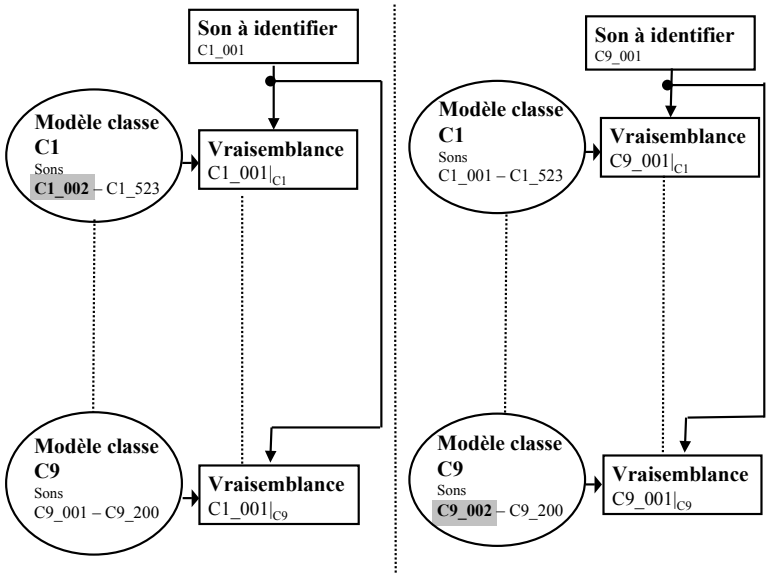


Fig. 4.13: Illustration du protocole de test «leave-one-out»

4.2.4.3 Choix du nombre de gaussiennes de la modélisation

Parmi les classes de sons il y a des classes qui peuvent contenir un nombre assez faible d'échantillons, l'utilisation d'un nombre trop important de gaussiennes ne conduira donc pas nécessairement à une meilleure reconnaissance. Le nombre d'échantillons de chaque classe a été donné dans le tableau 2.4 de la page 45. Pour que le modèle à base de gaussiennes (GMM) soit représentatif il faut que le nombre des vecteurs acoustiques de l'ensemble d'apprentissage soit suffisamment grand pour pouvoir estimer avec précision tous les paramètres des gaussiennes.

Parmi les critères de sélection de la meilleure modélisation d'une classe de son nous trouvons le critère BIC (Bayesian Information Criterion)[Schwarz, 1978] et AIC (Akaike Information Criterion) [Akaike, 1974]. La référence [Roeder and Wasserman, 1997] montre la fiabilité du critère BIC pour les mélanges des gaussiennes.

Avec ce critère la sélection du meilleur modèle est faite par la maximisation de la vraisemblance integrée :

$$(\widehat{m}, \widehat{K}) = \arg\max_{m,K} f(x \mid m, K) \tag{4.27}$$

où K est le nombre de composantes du modèle, m le modèle et $f(x \mid m, K)$ la vraisemblance intégrée telle que :

$$f(x \mid m, K) = \int_{\Theta_{m,K}} f(x \mid m, K, \theta)\pi(\theta \mid m, K)d\theta \tag{4.28}$$

$\Theta_{m,K}$ est l'espace des paramètres du modèle m avec K composantes. Une approximation asymptotique de la vraisemblance intégrée a été proposé par Schwarz [Schwarz, 1978] :

$$\log f(x \mid m, K) \approx \log f(x \mid m, K, \widehat{\theta}) - \frac{\nu_{m,K}}{2}\log(n) \tag{4.29}$$

où $\widehat{\theta}$ est l'estimation du θ et $\nu_{m,K}$ est le nombre de paramètres libres dans le modèle m. Le critère BIC est donné dans ce cas par l'équation :

$$BIC_{m,K} = -2.L_{m,K} + \nu_{m,K}\ln(n) \tag{4.30}$$

$L_{m,K}$ est le maximum logarithmique de la vraisemblance et il est égal à $\log f(x \mid m, K, \widehat{\theta})$. Pour trouver le modèle qui modélise aux mieux il faut minimiser la valeur BIC.

Comme la classe C9 (serrure de porte) a le nombre le plus réduit des trames (6050) disponible pour l'apprentissage, nous avons calculé le critère BIC pour cette classe pour 2,4,5 et 8 gaussiennes. Les calculs ont été effectués pour 16 coefficients MFCC. Les valeurs du critères BIC qui sont présentés dans le tableau 4.3 montrent qu'un nombre entre 3 et 5 gaussienne modélise le mieux cette classe.

Nombre de gaussiennes	2	**3**	**4**	**5**	8
BIC	11043	**10752**	**10743**	**10757**	13373

Tab. 4.3: BIC de la classe C9 pour 2, 3, 4, 5 et 8 gaussienne

4.2.4.4 Les résultats de la classification en utilisant un modèle GMM avec 4 gaussiennes

Une fois le nombre de gaussiennes pour le modèle GMM fixé à 4, les autres paramètres acoustiques sont testés par rapport à notre corpus (CPUR). Les paramètres utilisés sont : MFCC, LFCC, LPC, LPCC, fréquences Mel, bandes de fréquence rectangulaire, bandes de fréquence triangulaires. Des résultats avec la première et deuxième dérivée des paramètres MFCC sont présentés.

Le taux d'erreur de classification d'une classe est calculé comme suit :

$$\text{Taux d'erreur de classification} = \frac{\text{Nombre de sons non reconnus}}{\text{Nombre total de sons à reconnaître}} \tag{4.31}$$

Les résultats sont résumés dans le tableau 4.4 indiquant pour chaque combinaison des paramètres le taux moyen d'erreur de classification conformément à l'équation (4.31).

Type des paramètres	Nombre	Taux moyen d'erreur de classification [%]
16MFCC+Energie+Δ+$\Delta\Delta$	51	10.7
16MFCC+Energie+Δ	34	12.7
16LFCC+Energie	17	13.3
16MFCC+Energie	17	14.0
16LPCC	16	14.3
16MFCC	16	14.6
16LPC	16	20.1
3MFCC	3	21.7
16Mel	16	26.7
16 Rectangulaires	16	27.0
16 Triangulaires	16	32.0

Tab. 4.4: Taux moyen d'erreur de classification pour les paramètres classiques

Les meilleurs résultats sont obtenus avec les paramètres MFCC. La première dérivée améliore peu les résultats obtenus (1.3% - de 14% à 12.7%) par contre l'ajout de la deuxième dérivée diminue en absolu de 3.3% le taux d'erreur moyen (de 14% à 10.7%).

Les résultats de classification sont en concordance avec les valeurs du critère de Fisher. En tenant compte du critère de Fisher, la combinaison de trois coefficients MFCC (les coefficients 2, 3 et 4) a été testée. L'erreur augmente en absolu de 7% (de 14% à 21.7%) pour un nombre des paramètres 5 fois plus réduit.

Les paramètres LPC se fondent sur le modèle de la production de la parole, ce qui explique leurs performances médiocres sur des sons non-langagieres. La majorité des sons de la vie courante étudiés font partie des sons avec une répartition fréquentielle large, uniforme et de courte durée.

Les coefficients spectraux d'énergie ne caractérisent pas bien les sons de la vie courante, fait indiqué aussi par le critère de Fisher.

Pour une analyse de la confusion entre les différentes classes de sons nous pouvons nous rapporter à la matrice de confusion présentée dans le tableau 4.5 (elle a été obtenue pour 16 coefficients MFCC). Nous observons que pour la classe 7 (vaisselle) 58 sons sont reconnus comme appartenant à la classe 1 (bris de verre). Cela est dû à une forte ressemblance entre ces 2 types de sons, succession de bruits «percussifs».

	C1	C2	C3	C4	C5	C7	C9	Nbre total fichiers
C1 - Claquement de porte	**484**	3	0	36	0	0	0	523
C2 - Bris de verre	2	**85**	0	1	0	0	0	88
C3 - Sonneries de téléphone	23	0	**448**	0	45	1	0	517
C4 - Sons de pas	0	0	0	**13**	0	0	0	13
C5 - Cris	0	0	1	0	**72**	0	0	73
C7 - Vaisselle	58	12	4	0	0	**89**	0	163
C9 - Serrures de porte	4	2	0	0	0	0	**194**	200

Tab. 4.5: Matrice de confusion en nombre de sons pour un GMM avec 4 gaussiennes

4.3 Nouveaux paramètres acoustiques

Les paramètres acoustiques classiquement utilisés en parole donnent de bonnes performances (10% de taux moyen d'erreur dans le meilleur cas) de classification. Maintenant, nous recherchons de nouveaux paramètres acoustiques qui peuvent soit améliorer le taux moyen d'erreur de classification soit fournir les mêmes performances avec un nombre réduit de paramètres. Dans cette section, des paramètres utilisés plutôt en segmentation parole/musique/bruit sont présentés et évalués pour les sons de notre corpus (CPUR). Ensuite, de nouveaux paramètres issus de la transformée en ondelettes sont proposés et évalués.

4.3.1 ZCR - Le nombre de passages par zéro

Ce paramètre représente le nombre de passages par zéro du signal temporel dans la fenêtre d'analyse et généralement, il indique la fréquence dominante du signal dans la fenêtre.

Pour éliminer l'influence du bruit au voisinage de zéro (le bruit de quantification par exemple), on calcule le ZCR en utilisant un seuillage avec hystérésis comme montré dans la Figure 4.14. Dans notre application, nous utilisons un seuil égal à 10 pour des signaux variant

entre -32768 et 32767. La valeur du seuil représente 0.03% de la gamme d'entrée (10 fois la résolution de quantification) ; la valeur a été choisie en tenant compte des qualités de la carte d'acquisition.

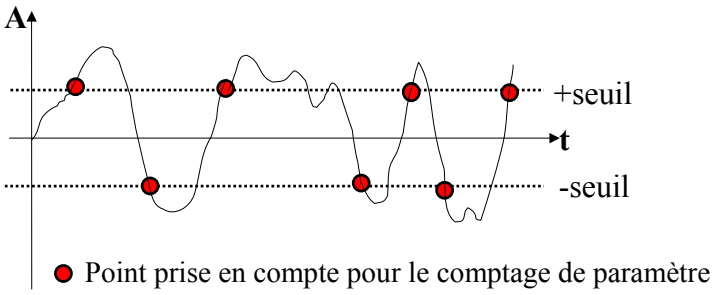


Fig. 4.14: Le nombre de passages par zéro

4.3.2 RF - Le Roll-off Point

RF est la fréquence au-dessous de laquelle se situe 95% de l'énergie du signal (Figure 4.15). On peut dire que c'est un indice de répartition du spectre de puissance du signal. Le Roll-off point est plus grand pour les signaux ayant un spectre de haute fréquence important. On le calcule suivant la formule (4.32) avec $\Upsilon = 0.95$.

$$RF = \alpha \quad \text{avec } \alpha \text{ tel que} \quad \sum_{k<\alpha} X[k] = \Upsilon \sum_{k} X[k] \tag{4.32}$$

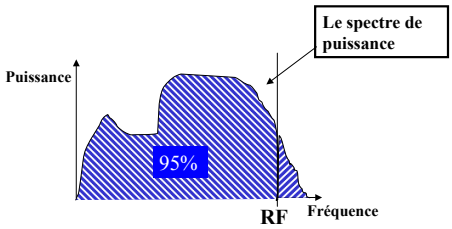


Fig. 4.15: Roll-off point sur le spectre de puissance du signal

4.3.3 Le Centroïde Spectral

Le centroïde spectral est la valeur de la fréquence partageant le spectre en deux parties d'égale énergie : basse fréquence/haute fréquence (Figure 4.16). Son calcul se fait avec la même formule que le Roll-off point (4.32) où $\Upsilon = 0.5$.

Le spectre du signal est scindé en K bandes de fréquence de largeur fixée, $X[k]$ est la puissance du signal de la bande de fréquence d'indice k.

4.3.4 Coefficients provenant de la transformée en ondelettes

Les coefficients proposés dans cette section proviennent de la transformée en ondelettes. Comme nous l'avons vu dans la section 3.3.3.1 (page 80) la transformée en ondelettes est

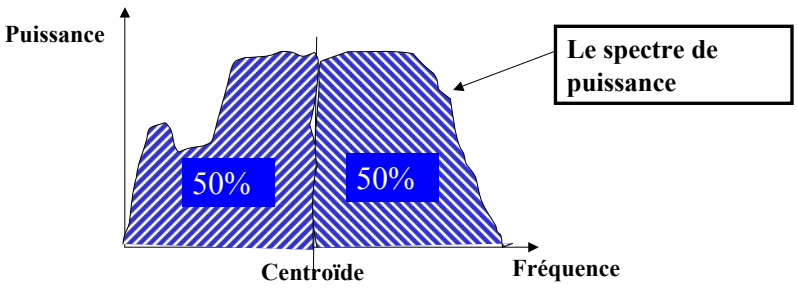


Fig. 4.16: Centroïde sur le spectre de puissance du signal

mieux adaptée à l'analyse des signaux impulsionnels parce que sa résolution fréquencielle diffère d'une bande de fréquences à l'autre.

La répartition des coefficients de la transformée en ondelettes sur 256 échantillons est celle présentée dans la Figure 3.29 de la page 84. La même durée de la fenêtre d'analyse que celle utilisée pour les paramètres acoustiques classiques a été choisie (de même pour la superposition des fenêtres d'analyse). La taille des coefficients de la transformée en ondelettes est très variable, commençant avec un échantillon pour les deux premiers coefficients, en doublant chaque fois pour arriver à une taille de 128 échantillons pour le neuvième et dernier coefficient [Truchetet, 1998].

Les coefficients de la transformée en ondelettes $C_1, \ldots, C_9$ représentent la décomposition en fréquence et en temps du signal par les fonctions de type ondelettes [Mallat, 2000].

La transformée en ondelettes d'un claquement de porte (Figure 4.17) semble contenir une information spectrale concentrée dans les hautes fréquences, respectivement les coefficients de C_5 à C_9.

En tenant compte de la structure des coefficients de la transformée en ondelettes et du fait qu'on n'a pas défini dans la littérature jusqu'à maintenant des paramètres acoustiques issus de cette transformée, une étude exploratoire a été effectuée. Parmi les caractéristiques des coefficients en ondelettes testés comme paramètres acoustiques, les trois meilleurs sont présentés.

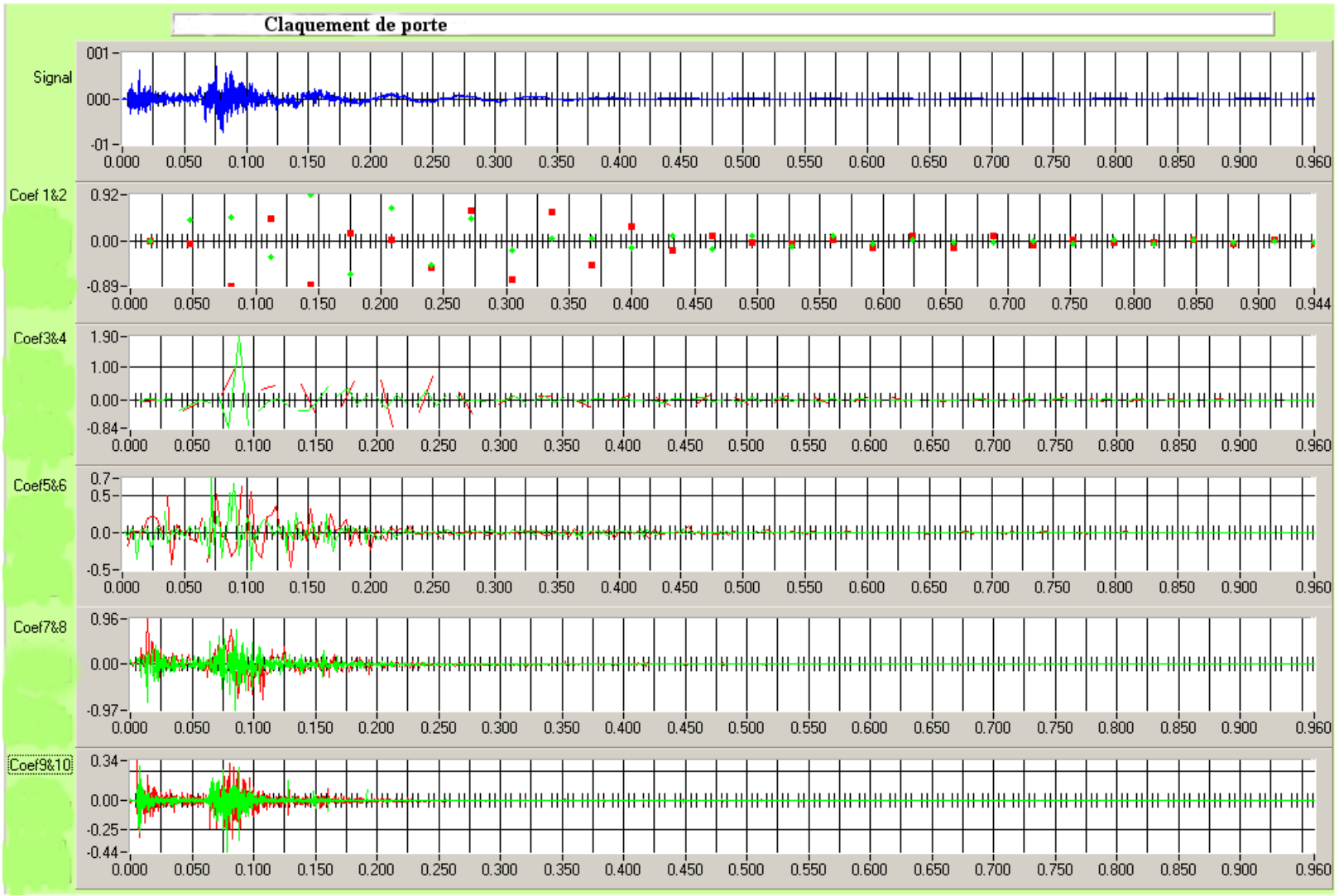


Fig. 4.17: Transformée en ondelettes du son d'un claquement de porte

4.3.4.1 La moyenne et l'écart-type des coefficients de la transformée en ondelettes

Les premières caractéristiques des coefficients choisis comme paramètres acoustiques sont la moyenne (μ) et l'écart-type (σ). Un vecteur acoustique de 10 éléments, représenté dans la Figure 4.18 a été utilisé. Seulement la moyenne et l'écart-type des 5 derniers coefficients ont été pris en compte parce que les premiers 4 coefficients ont très peu d'échantillons (1, 1, 2 et respectivement 4).

μ_{C_5}	σ_{C_5}	μ_{C_6}	σ_{C_6}	μ_{C_7}	σ_{C_7}	μ_{C_8}	σ_{C_8}	μ_{C_9}	σ_{C_9}

Fig. 4.18: Vecteur acoustique constitué de la moyenne et l'écart-type des 5 derniers coefficients de la transformée en ondelettes (taille de la fenêtre 16 ms avec recouvrement de 50%)

4.3.4.2 L'écart-type, l'énergie, et les moments d'ordre supérieur (Skewness et Kurtosis)

Le deuxième vecteur acoustique proposé est composé de l'écart-type (σ), le Skewness (S), le Kurtosis (K) et l'énergie (E) des 5 derniers coefficients de la transformée en ondelettes [Lacoume, 1997]. La dimension du vecteur est 20, comme le représente la Figure 4.19. Les mêmes raisons qu'avant, déterminent le choix des 5 derniers coefficients seulement de la transformée en ondelettes.

σ_{C_5}	S_{C_5}	K_{C_5}	E_{C_5}	$\ldots$	σ_{C_9}	S_{C_9}	K_{C_9}	E_{C_9}

Fig. 4.19: Vecteur acoustique fondé sur l'écart-type, le Skewness, le Kurtosis et l'énergie des 5 derniers coefficients de la transformée en ondelettes

La formule du calcul du $Skewness$ (statistique d'ordre 3) est (4.33) où σ est l'écart-type (équation (4.34)), s_i l'échantillon i du signal et $\overline{s}$ est la moyenne du signal.

$$Skewness = \frac{1}{N}\sum_{i=0}^{N-1}\left(\frac{s_i - \overline{s}}{\sigma}\right)^3 \tag{4.33}$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=0}^{N-1}(s_i - \overline{s})^2} \tag{4.34}$$

La formule du $Kurtosis$ (statistique d'ordre 4) est (4.35) où σ est l'écart-type (équation (4.34)).

$$Kurtosis = \left[\frac{1}{N}\sum_{i=0}^{N-1}\left(\frac{s_i - \overline{s}}{\sigma}\right)^4\right] - 3 \tag{4.35}$$

4.3.4.3 DWTC - Les coefficients «cepstraux» de la transformée en ondelettes

Un autre type de paramètres proposés est issu de la transformée en ondelettes et est fondé sur le principe des coefficients cepstraux, conformément à la Figure 4.20. Premièrement, le calcul de la transformée en ondelettes sur 256 échantillons (X_i) est effectué ($N = 256$), donc $\widetilde{X}_i$sont

obtenus. Ensuite, l'énergie des 6 derniers coefficients ($M = 6$) de la transformée est calculée (E_i), suivie de l'application du logarithme décimal ($\widehat{E}_i$). Le calcul du vecteur acoustique se termine par le calcul de la transformée en ondelettes inverse du vecteur d'énergies logarithmique (P_i).

La dimension du vecteur acoustique est de 6 éléments.

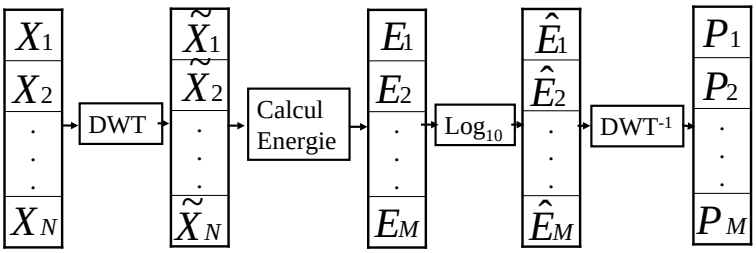


Fig. 4.20: Principe de coefficients «cepstraux» issus de la transformée en ondelettes

4.3.5 Statistiques des nouveaux paramètres acoustiques

L'étude de la pertinence des paramètres acoustiques a été effectuée en calculant le critère de Fisher FDR (Fisher Discriminant Ratio) pour tous les nouveaux paramètres.

4.3.5.1 Le nombre de passages par zéro, le centroïde et le roll-off point

Les valeurs du critère de Fisher se trouvent dans le tableau 4.6 pour les paramètres acoustiques suivants : le nombre de passages par zéro, le centroïde et le roll-off point. Les valeurs élevées du critère de Fisher montrent une bonne pertinence de ces paramètres.

Paramètre	FDR
Nombre de passages par zéro	18
Centroïde	23.75
Roll-off point	16.70

Tab. 4.6: Valeurs du rapport de Fisher pour le nombre de passages par zéro, le centroïde et le roll-off point

4.3.5.2 Les paramètres issus de la transformée en ondelettes

Les valeurs du critère de Fisher pour les paramètres issus de la transformée en ondelettes se trouvent dans le tableau 4.7. La moyenne des coefficients en ondelettes ne peut pas être utilisée parce que les coefficients sont centrés sur zéro. Par contre l'écart-type des coefficients de C_6 à C_9 (ceux qui correspondent aux hautes fréquences) semblent potentiellement intéressants. L'énergie des coefficients a une valeur de FDR plus grande que 2, seulement pour les coefficients de C_7 à C_9. De même pour le skewness et le kurtosis.

Les DWTC, les «cepstres» des coefficients en ondelettes sont les mieux classés par le critère de Fisher parmi les coefficients issus de la transformée en ondelettes.

Paramètre	FDR	Paramètre	FDR	Paramètre	FDR
μ_{C_5}	0.01	σ_{C_5}	0.28	$DWTC_1$	1.84
σ_{C_5}	0.81	S_{C_5}	0.32	$\mathbf{DWTC_2}$	**1.92**
μ_{C_6}	0	K_{C_5}	0.53	$\mathbf{DWTC_3}$	**2.89**
$\sigma_{\mathbf{C_6}}$	**1.98**	E_{C_5}	0.24	$\mathbf{DWTC_4}$	**4.54**
μ_{C_7}	0	σ_{C_6}	0.38	$\mathbf{DWTC_5}$	**6.02**
$\sigma_{\mathbf{C_7}}$	**4.64**	S_{C_6}	0.97	$\mathbf{DWTC_6}$	**8.69**
μ_{C_8}	0.01	K_{C_6}	1.26	-	-
$\sigma_{\mathbf{C_8}}$	**3.82**	E_{C_6}	0.52	-	-
μ_{C_9}	0.01	σ_{C_7}	1.33	-	-
$\sigma_{\mathbf{C_9}}$	**6.74**	$\mathbf{S_{C_7}}$	**2.01**	-	-
-	-	$\mathbf{K_{C_7}}$	**2.68**	-	-
-	-	$\mathbf{E_{C_7}}$	**2.10**	-	-
-	-	$\sigma_{\mathbf{C_8}}$	**3.11**	-	-
-	-	$\mathbf{S_{C_8}}$	**5.17**	-	-
-	-	$\mathbf{K_{C_8}}$	**6.10**	-	-
-	-	$\mathbf{E_{C_8}}$	**4.54**	-	-
-	-	$\sigma_{\mathbf{C_9}}$	**3.04**	-	-
-	-	$\mathbf{S_{C_9}}$	**4.24**	-	-
-	-	$\mathbf{K_{C_9}}$	**4.94**	-	-
-	-	$\mathbf{E_{C_9}}$	**4.13**	-	-

Tab. 4.7: Valeurs du rapport de Fisher pour les paramètres issus de la transformée en ondelettes

Nous pouvons observer que les paramètres issus de la transformée en ondelettes présentent de faibles valeurs du FDR en rapport avec les paramètres classiques, nous allons cependant les tester dans la tâche de classification.

4.3.6 Résultats de la classification avec les nouveaux paramètres acoustiques

Après l'étude statistique présentée ci-dessus, ces nouveaux paramètres acoustiques ont été testés sur notre corpus de sons (CPUR). Les combinaisons entre les trois paramètres non classiques en reconnaissance de parole et les MFCC ont été validés. Les résultats se trouvent dans le tableau 4.8.

Dans le tableau 4.8 nous avons groupé les performances des nouveaux paramètres acoustiques avec les paramètres classiques.

Au niveau des paramètres issus de la transformée en ondelettes en tenant compte des valeurs du critère de Fisher les combinaisons suivantes ont été testées :

- 4 valeurs d'écart-type (du coefficient 6 à 9)
- 20 valeurs d'écart-type, énergie, skewness et kurtosis
- les «cepstres» des 6 coefficients en ondelettes (du coefficient 4 à 9)(DWTC)
- 4 valeurs d'écart-type (du coefficient 6 à 9) combinés avec le Roll-off point, le nombre de passage par zéro et le Centroïde

- 4 valeurs d'écart-type (du coefficient 6 à 9) combinés avec 16MFCC en conjonction avec le nombre de passages par zéro, le Roll-off Point, le Centroïde et l'énergie

Seuls les trois paramètres non classiques (le nombre de passages par zéro, le roll-off point et le centroïde)(18) conduisent à un taux d'erreur de classification de 24.6% qui est meilleur que celui obtenu avec les coefficients d'énergie spectraux (voir tableau 4.4 de la page 116). Par contre, en rajoutant à ces paramètres les trois plus pertinents coefficients MFCC (9), une amélioration en absolu des performances de 8% est obtenue (de 24.6% à 16.3%). Si on rajoute aux trois paramètres tous les 16 coefficients MFCC (5) une amélioration absolue de 14.5% est obtenue (de 24.6% à 10.1%).

La dérivée première des coefficients n'améliore pas la classification obtenue par la combinaisons des MFCC avec les trois paramètres non classiques (3). Par contre la deuxième dérivée (1) améliore en absolu de 3% les performances (de 10.1% à 7.1%).

Nr.	Type des paramètres	Nombre	Taux moyen d'erreur de classification [%]
1	16MFCC+Énergie+ZCR+RF+Centroïde+Δ+ΔΔ	60	7.1
2	*16MFCC+Énergie*+Δ+ΔΔ	*51*	*10.7*
3	16MFCC+Énergie+ZCR+RF+Centroïde+Δ	40	10.0
4	*16MFCC+Énergie*+Δ	*34*	*12.7*
5	16MFCC+Énergie+ZCR+RF+Centroïde	20	10.1
6	*16MFCC+Énergie*	*17*	*14.0*
7	16LFCC+ZCR+RF+Centroïde	19	12.7
8	*16LFCC+E*	*17*	*13.3*
9	3MFCC+ZCR+RF+Centroïde	6	16.3
10	*3MFCC*	*3*	*21.7*
11	16MFCC+ZCR+RF+Centroïde+Énergie+ +4ondelettes(écart-type)	24	16.5
12	16LPCC+ZCR+RF+Centroïde	19	17.4
13	*16LPCC*	*16*	*14.3*
14	4ondelettes(écart-type)	4	18.6
15	4ondelettes(écart-type)+ZCR+RF+Centroïde	7	18.3
16	6ondelettes(DWTC)	6	18.7
17	20ondelettes(5écarts-types+5énergies+ +5skewness+5kurtosis)	20	20.3
18	ZCR+RF+Centroïde	3	24.6

Tab. 4.8: Taux moyen d'erreur de classification pour les nouveaux paramètres

Parmi les paramètres issus de la transformée en ondelettes, les DWTC (16) et ceux fondés sur l'écart-type (14) donnent un taux d'erreur de classification approximatif de 18.6% qui est un taux de 8% plus grand que celui des MFCC combinés avec les trois nouveaux paramètres (5). Cependant, les DWTC sont réduits à seulement 6 paramètres au lieu de 20. Par la suite, d'autres propriétés de ces coefficients seront montrées.

L'addition du Roll-off point, du nombre de passages par zéro et du Centroïde aux meilleurs paramètres issus de la transformée en ondelettes (4 écarts-types) (15) n'améliore pas les performances de ceux-ci.

L'addition des paramètres issus de la transformée en ondelettes aux 16MFCC couplés avec le nombre de passages par zéro, le Roll-off Point, le Centroïde et l'énergie (11) n'améliore pas les performances qui peut s'expliquer par la redondance des informations spectrales contenues d'une part dans les MFCC et d'autre part dans les paramètres issus de la transformée en ondelettes.

4.3.7 Apport des nouveaux paramètres à la classification des sons de la vie courante

Après avoir étudié la pertinence statistique des paramètres acoustiques classiques et non classiques pour le corpus des sons de la vie courante nous pouvons constater que le meilleur taux d'erreur de classification obtenu a été de 7.16%. Ce taux est obtenu pour la deuxième et la première dérivée additionnées à la combinaison des 16 MFCC avec les trois paramètres non classiques (ZCR, RF et le Centroïde). Cette combinaison totalise 60 paramètres acoustiques ce qui tend à augmenter le temps de calcul.

Le meilleur compromis performances / nombre de paramètres a été obtenu pour la combinaison des 16 MFCC avec les trois paramètres non classiques qui donne un taux d'erreur de 10.15%.

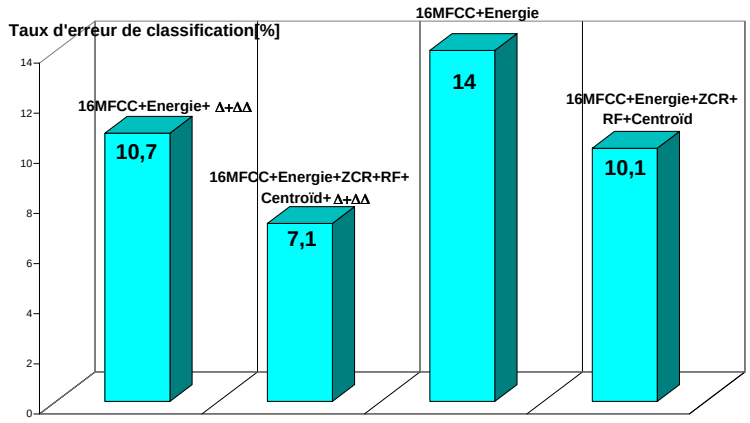


Fig. 4.21: Amélioration des performances de la classification par combinaison entre les MFCC et les trois paramètres non classiques

La Figure 4.21 montre que l'addition des paramètres non classiques aux coefficients MFCC améliore les performances de classification. Dans le cas des 16 MFCC par rajout des nouveaux paramètres, le taux d'erreur de classification diminue de 14% à 10.1% et dans le cas de l'addition de la deuxième et première dérivée avec les mêmes coefficients le taux d'erreur diminue de 10.7% à 7.16%. En conclusion, les paramètres acoustiques non classiques permettent une amélioration significative des performances des coefficients MFCC (les plus pertinents paramètres

acoustiques classiques).

Les paramètres issus de la transformée en ondelettes ont quant à eux un taux d'erreur de classification de 18% pour seulement 6 coefficients.

4.4 Classification de sons bruités

Après avoir validé l'algorithme de classification à base de modèles GMM et avoir trouvé les paramètres acoustiques les plus pertinents pour le type de sons concerné, l'étape suivante est d'étudier le problème des sons bruités.

Premièrement, une étude avec les paramètres acoustiques les plus pertinents a été effectuée sur le corpus des classes de sons bruités (CBRUIT) décrits dans la section 2.3.2 de la page 45. Le but de cette étude est de trouver les paramètres acoustiques les plus robustes au bruit.

L'approche pour améliorer les performances de classification dans le bruit est celle du débruitage avant la phase de classification. Une méthode à base d'ondelettes a été étudiée.

4.4.1 Étude de la résistance des paramètres acoustiques au bruit

Pour étudier le comportement du système de classification en présence de bruit une base de test bruitée avec le son de l'appartement HIS a été créée, conformément à la section 2.3.2 (page 45). Les sons purs sont considérés comme ayant un RSB de 70 dB.

Le protocole de test utilisé est le même que celui présenté dans la section 4.2.4.2 (page 114) avec les différences suivantes :

- nous testons seulement les sons bruités (avec un RSB de 0, 10, 20 et 40 dB) et non pas les sons purs
- la base d'apprentissage est constituée seulement des sons purs
- quand nous testons un son, nous éliminons de la base d'apprentissage sa variante pure

Nous calculons un taux moyen d'erreur de classification pour chaque classe de son et pour chaque RSB. La formule de calcul est la même (équation (4.31) de la page 116).

Les coefficients MFCC avec soustraction de la moyenne ne sont pas étudiés parce que dans notre cas le bruit est additif et non pas convolutif. L'hypothèse de départ de ces paramètres est que le bruit est rajouté aux sons par une convolution dans le temps.

4.4.1.1 Résultats obtenus

La dépendance du taux moyen d'erreur de classification en fonction du RSB des sons (0, 10, 20, 40 dB et 70dB pour les sons purs) est donnée dans la Figure 4.22 pour des paramètres du type MFCC en conjonction avec le Roll-off point, le nombre de passages par zéro, le Centroïd et l'énergie, LFCC, LPCC et des paramètres issus de la transformée en ondelettes.

Les paramètres issus de la transformée en ondelettes procurent les meilleures performances dans le cas où le RSB est inférieur à 20dB (amélioration en absolu de 8%), mais, ces résultats demeurent inférieurs de 10% à ceux obtenus par les MFCC combinés avec les trois paramètres non classiques en reconnaissance de la parole lorsque le RSB est supérieur à 20 dB.

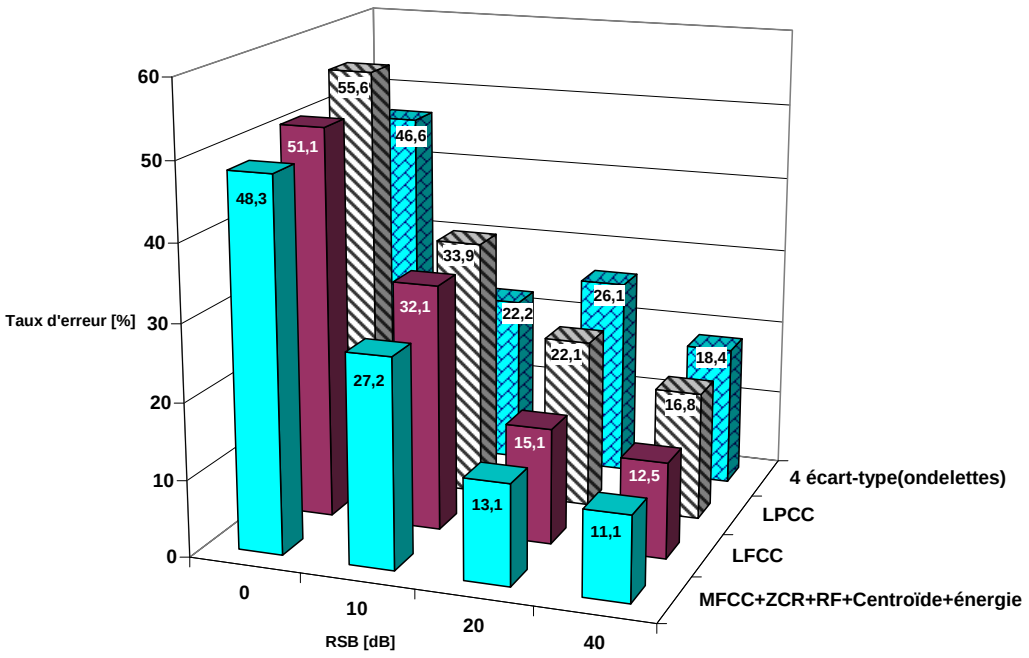


Fig. 4.22: Variation du taux moyen d'erreur de classification en fonction du RSB pour les coefficients MFCC, LFCC, LPCC et ceux issus de la transformée en ondelettes

4.4.2 Débruitage des sons avant la phase de classification avec la transformée en ondelettes

Cette approche consiste en un filtrage du signal avant le calcul des paramètres acoustiques et de la classification. Le vaste domaine de débruitage des signaux comprend des méthodes comme le filtrage de Wiener, le filtrage par ondelettes [G. Tzanetakis and Cook, 2001], le filtrage adaptatif, la soustraction spectrale, etc.

La transformée en ondelettes est mieux adaptée à l'analyse et au traitement des signaux impulsionnels par rapport à la transformée de Fourier qui est adaptée aux signaux périodiques (voir section 3.3.3 - page 80). Tenant compte des propriétés de la transformée en ondelettes nous avons choisi de l'utiliser pour le débruitage des signaux avant la phase de classification [H. Hacihabiboglu and Nishan, 2002]. Les ondelettes sont de plus en plus utilisées pour le filtrage des signaux sonores [Antoniadis, 2003], [Mallat, 2000].

L'algorithme utilisé comprend les étapes suivantes, conformément au schéma de la Figure 4.23 :

- Calcul de la transformée en ondelettes (DWT) sur une fenêtre de 256 échantillons
- Seuillage des coefficients de la transformée en ondelettes
- Calcul de la transformée en ondelettes inverse (DWT^{-1})

Les seuils sont appliqués sur le module de chaque coefficient de la transformée en ondelettes. Les valeurs des seuils sont estimées au lancement de l'algorithme sur les premières 100 ms de signal qui sont considérées comme contenant seulement le bruit environnemental et non

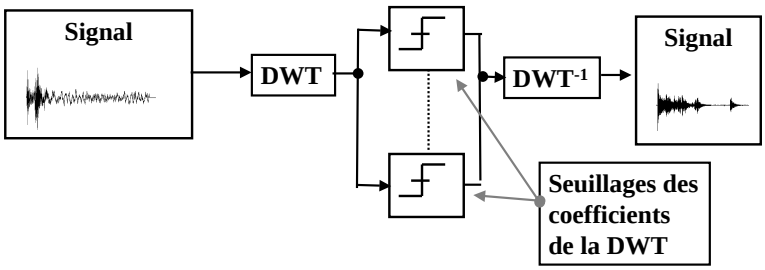


Fig. 4.23: Schéma de l'algorithme de débruitage avec la transformée en ondelettes

pas le signal. En conclusion les valeurs des seuils sont estimés pour chaque événement sonore à identifier. Les valeurs de seuils sont les suivantes :

$$\begin{cases} S_i = 1.2 * B^i_{max} & \text{pour} \quad i = 1 \dots 4 \\ S_i = 0.9 * B^i_{max} & \text{pour} \quad i = 5 \\ S_i = 0 & \text{pour} \quad i = 6 \dots 9 \end{cases}$$

où :

$$\begin{cases} S_i & \text{est le seuil appliqué au coefficient } i \text{ de la transformée en ondelettes} \\ B^i_{max} & \text{est l'amplitude maximale du coefficient } i \text{ de la transformée en ondelettes pour le} \\ & \text{bruit. Cette valeur est calculée pendant les 100 premières millisecondes du signal.} \end{cases}$$

Remarque : Une valeur égale à 0 pour un seuil signifie l'absence du filtrage dans la bande fréquence correspondante.

Ce choix des seuils de filtrage a été fait en étudiant les coefficients de la transformée en ondelettes pour le bruit HIS et pour les sons à identifier. Les sons à identifier n'ont pas d'information utile dans les 5 premiers coefficients, par contre le bruit HIS est concentré dans les basses fréquences.

Un exemple d'un son de bris de verre mélangé avec le bruit de l'appartement HIS à un RSB de 0 dB filtré par la méthode ci-dessus, est donné dans la figure 4.24. Dans la première fenêtre est illustré le signal bruité et dans la deuxième le signal après filtrage par ondelettes.

Les résultats de classification obtenus avec 16 paramètres MFCC combinés avec le Roll-off point, le nombre de passages par zéro, le Centroïde et l'énergie comme paramètres acoustiques, sur les signaux filtrés avec cette méthode sont présentés dans le tableau 4.9. Nous observons qu'avec ce filtrage pour des RSB plus petits ou égaux à 10 dB nous avons un gain absolu au niveau du taux moyen d'erreur de classification de 8%. Pour un RSB égal à 20dB le gain est négligeable et pour un RSB égal à 40dB le taux devient égal à celui pour les sons purs (RSB=70dB).

4.5 Conclusions

Dans ce chapitre la classification des sons de la vie courante a été abordée. La première approche de classification des sons purs consiste à utiliser un système à base de GMM et des paramètres acoustiques classiquement exploités dans la reconnaissance de la parole/locuteur.

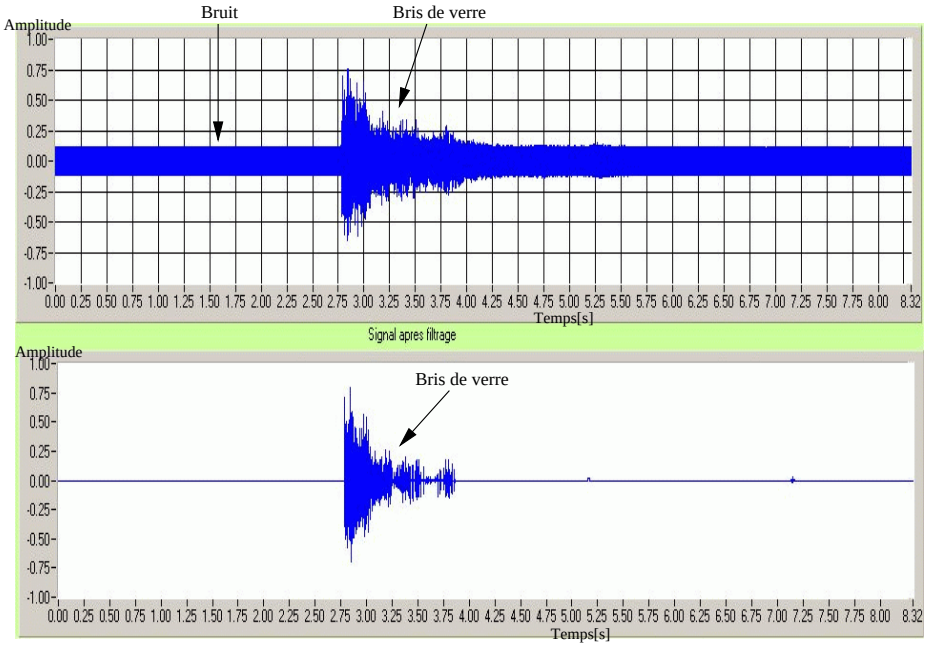


Fig. 4.24: Son de verre cassé avec du bruit de l'appartement HIS avec un RSB= 0dB avant et après le filtrage

Taux d'erreur / Avec/Sans filtrage	RSB			
	0 dB	10 dB	20 dB	40 dB
Avec	40.0	20.5	14.6	11
Sans	48.3	27.2	13.1	11

Tab. 4.9: Taux moyen d'erreur de classification par RSB avec 16MFCC en conjonction avec le nombre de passages par zéro, le Roll-off Point, le Centroïd et l'énergie avec et sans filtrage par transformée en ondelettes

Des paramètres non classiques et de nouveaux paramètres issus de la transformée en ondelettes ont été ensuite validés.

Le système à base de GMM ne tient pas compte de l'évolution temporelle du signal et donc de sa durée ce qui implique la nécessité pour la phase d'apprentissage d'un grand nombre de sons pour pallier leur durée réduite (petit nombre de trames acoustiques). L'ajout d'informations temporelle avec l'utilisation de la première et de la deuxième dérivée permet un gain des performances.

Le couplage entre les paramètres classiques de type MFCC et le nombre de passages par zéro, le Roll-off point et le centroïde donne les meilleures performances. L'étude statistique effectuée nous a permis d'atteindre une réduction importante du nombre des paramètres acoustiques avec une diminution des performances acceptable.

Ensuite une étude de l'influence du bruit sur le système de classification des sons a été effectuée. Les coefficients issus de la transformée en ondelettes semblent être «plus résistants» au bruit avec un RSB faible (RSB plus petit ou égal à 10 dB).

Une solution d'amélioration des performances dans le bruit a été proposée : le filtrage par seuillage des coefficients de la transformée en ondelettes. Ce filtrage apporte un gain non négli-

geable de 8% sur le taux moyen d'erreur de classification.

Pour la classification, les contraintes «temps réel» sont moindres parce que ce processus tourne en temps différé. Une mesure des temps de calcul de la classification a cependant été effectuée. Le temps de classification d'un son d'une longueur de 7 secondes (longueur maximale obtenue après la détection) avec 6 paramètres acoustiques est de 0.5 secondes et avec 60 paramètres acoustiques est d'une seconde. Dans le cas des 6 paramètres avec débruitage pour un son de 7 secondes nous avons un temps de calcul d'une seconde et pour 60 paramètres le temps de calcul est de 2 secondes.

En conclusion, la classification des sons de la vie courante a une erreur moyenne sur les sons purs de 10% et sur les sons bruités à un RSB de 10 dB de 22% pour notre application. En réalité les sons avec un RSB plus petit que 20 dB sont rares. Dans ce cas, un taux d'erreur de classification moyen de 10% en conjonction avec la fusion de données qui a lieu dans le cadre du système de télésurveillance médical confirment l'utilisabilité du système proposé.

Bibliographie

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automation Control*, 19 :716–723.

[Antoniadis, 2003] Antoniadis, A. (2003). Compression et débruitage avec les ondelettes. Technical report, LMC - IMAG.

[Boite et al., 2000] Boite, R., Bourlard, H., Dutoit, T., Hang, J., and Leich, H. (2000). *Traitement de la parole*. ISBN 2-88074-388-5. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Cabell and Fuller, 1989] Cabell, R. H. and Fuller, C. R. (1989). A smart pattern recognition system for the automatic identification of aerospace acoustic sources. *AIAA 12th Aeroacoustics Conference*, pages 321–325, San Antonio, USA.

[CALLIOPE, 1989] CALLIOPE (1989). *La parole et son traitement automatique*. ISBN 2-225-81516-X. Masson, Paris.

[Cappé, 2001] Cappé, O. (2001). Modèles de mélange et modèles de markov cachés pour le traitement automatique de la parole. *[http ://tsi.enst.fr/ cappe/h2m/index.html]*, (ENST/Paris) :1–9.

[Chagnolleau et al., 2001] Chagnolleau, I. M., Gravier, G., and Blouet, R. (2001). Overview of the ELISA consortium research activities. *2001 : a Speaker Odyssey*, (2) :67–72.

[Couvreur, 1997] Couvreur, C. (1997). *Environmental Sound Recognition : A statistical approach*. PhD thesis, Faculté Polytechnique de Mons, Belgique.

[Cowling and Sitte, 2002] Cowling, M. and Sitte, R. (2002). Analysis of non speech recognition techniques for use in a non-speech sound recognition system. *6th International Symposium on Digital Signal Processing for Communication Systems - Sydney-Manly, Australie*, page 4pages.

[Davis and Mermelstein, 1980] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Audio, Speech and Signal Processing*, 28(4) :357–366.

[Dufaux, 2001] Dufaux, A. (2001). *Detection and recognition of Impulsive Sounds Signals*. PhD thesis, Faculté des sciences de l'Université de Neuchâtel, Suisse.

[El-Maleh et al., 2000] El-Maleh, K., Klein, M., Petrucci, G., and Kubal, P. (2000). Speech/music discrimination for multimedia applications. *International Conference on Acoustics, Sound and Signal Processing (ICASSP), Istanbul, République Turque*.

[G. Tzanetakis and Cook, 2001] G. Tzanetakis, G. E. and Cook, P. (2001). Audio analysis using the discrete wavelet transform. *WSES Int. Conf. Acoustics and Music : Theory and Applications (AMTA 2001)*, pages 225–229, Skiathos, Greece.

[Goldhor, 1993] Goldhor, R. S. (1993). Recognition of environmental sounds. *International Conference on Audio, Speech and Signal Processing*, volume 1, pages 149–152, Minneapolis, USA.

[Goldstein, 1976] Goldstein, U. (1976). Speaker-identifying features based on formant tracks. *Journal of the Acoustical Society of America*, 59(1) :176–182.

[H. Hacihabiboglu and Nishan, 2002] H. Hacihabiboglu, F. H. and Nishan, C. (2002). Musical instrument recognition with wavelet envelopes. *EAA Convention, Proc. Forum Acusticum Sevilla*, pages 356–360, Sevilla, Spain.

[Häcker et al., 2002] Häcker, J., Engelhardt, F., and Frey, D. D. (2002). Robust manufacturing inspection and classification with machine vision. *International Journal of Production Research*, 40(6) :1319–1334.

[JAIN et al., 2000] JAIN, A. K., DIUN, R. P. W., and MOA, J. (2000). Statistical pattern recognition : A review. *IEEE Trans. PAMI*, 22(1) :4–37.

[Jolliffe, 1986] Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

[Jouvet, 1988] Jouvet, D. (1988). *Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris.

[Kil and Shin, 1996] Kil, D. and Shin, F. (1996). *Pattern Recognition and Prediction with Applications to Signal Characterization*. AIP Press Woodbury, New York.

[Kunt et al., 2000] Kunt, M., Coray, G., Granlund, G., Haton, J., Ingold, R., and Kocher, M. (2000). *Reconnaissance des formes et analyse de scènes*, volume 1-3 of *ISBN 2-88074-384-2*. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Lacoume, 1997] Lacoume, J. L. (1997). *Statistiques d'ordre supérieur pour le traitement du signal*. ISBN 2-225-83118-1. Masson, Paris.

[Linde et al., 1980] Linde, Y., Buzo, A., and Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, COM-28 :84–95.

[Mallat, 2000] Mallat, S. (2000). *Une exploration des signaux en ondelettes*. ISBN 2-7302-0733-3. Les Editions de l'Ecole Polytechnique, Paris.

[Myers and Rabiner, 1981] Myers, C. S. and Rabiner, L. R. (1981). A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7) :1389–1409.

[Papadopoulos et al., 1992] Papadopoulos, G., Efstathiou, K., Li, Y., and Delis, A. (1992). Implementation of an intelligent instrument for passive recognition and two-dimensional location estimation of acoustic targets. *IEEE Transactions On Instrumentation and Measurement*, 41(6) :885–890.

[Paradie and Nawab, 1990] Paradie, M. J. and Nawab, S. H. (1990). The classification of ringing sounds. *International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 2435–2438, New Mexico, USA.

[Pfeiffer et al., 1996] Pfeiffer, S., Fischer, S., and Effelsberg, W. (1996). Automatic audio content analysis. *Reihe Informatik*, 8 :15–27.

[Rabiner and Juang, 1993] Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of speech recognition*. ISBN 0-13-015157-2. Prentice Hall PTR, New Jersey, USA.

[Reynolds, 1994] Reynolds, D. (1994). Speaker identification and verification using gaussian mixture speaker models. *Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Suisse*, pages 27–30.

[Roeder and Wasserman, 1997] Roeder, K. and Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92 :894–902.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464.

[Scott et al., 1993] Scott, E. A., Fuller, C. R., O'Brien, W. F., and Cabell, R. H. (1993). Sparse distributed associative memory for the identification of aerospace acoustic sources. *AIAA Journal*, 31(9) :1583–1589.

[Seck et al., 2001] Seck, M., Magrin-Chagnolleau, I., and Bimbot, F. (2001). Experiments on speech tracking in audio documents using gaussian mixture modeling. *International Conference on Acoustics, Sounds and Signal Processing*, Salt Lake City, Utah, USA.

[Theodoridis and Koutroumbas, 1998] Theodoridis, S. and Koutroumbas, K. (1998). *Pattern Recognition*. Academic Press, San Diego.

[Truchetet, 1998] Truchetet, F. (1998). *Ondelettes pour le signal numérique*. ISBN 2-86601-672-6. Hermes.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, IT-13 :260–269.

[Woodard, 1992] Woodard, J. P. (1992). Modeling and classification of natural sounds by product code hidden markov models. *IEEE Transactions on Signal Processing*, 40(7) :1833–1835.

[Yaniv and Burshtein, 2003] Yaniv, R. and Burshtein, D. (2003). An enhanced dynamic time warping model for improved estimation of DTW parameters. *IEEE Transactions on Speech and Audio Processing*, 11(3) :216–228.

Chapitre 5

Couplage entre détection et classification

Ce chapitre présente les problématiques du couplage entre le système de détection des événements sonores et celui de classification des sons. Les blocs fonctionnels considérés dans ce chapitre sont présentées en gras dans la Figure 5.1. La détection d'un évènement sonore consiste à déterminer l'instant d'apparition et de détection d'un signal, en suite de quoi le système de classification est activé en vue d'identifier cet évènement.

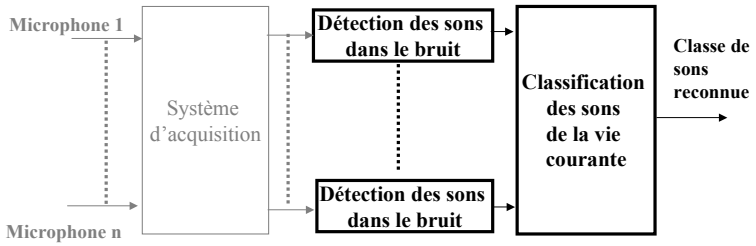


Fig. 5.1: Structure du système d'extraction d'informations sonores

Le premier point - critique - affectant la qualité du couplage entre classification et détection consiste à bien déterminer le début et la fin du signal détecté avant l'activation du système de classification des sons. Parmi les erreurs possibles nous avons :

- Détection prématurée du signal, auquel cas une partie du signal soumis à la classification ne contient que du bruit ;
- Détection retardée du signal, auquel cas une partie du signal utile n'est pas pris en compte, ce qui peut affecter la qualité de la classification. La classification étant cependant effectuée avec un système de GMM qui ne tient pas compte de l'évolution temporelle du signal, l'influence de cette erreur ne devrait pas être dramatique ;
- une fausse alarme au niveau de la détection reste une fausse alarme pour le système complet.

L'algorithme de détection fondé sur la transformée en ondelettes, proposé dans le chapitre 3, a les meilleures performances de détection et une bonne précision de détermination de l'instant du début du signal. Dans [Dufaux, 2001], une méthode de seuillage appliquée pendant les 1.5 secondes précédant la détection du signal est proposée pour améliorer la précision.

Le deuxième point affectant la qualité du couplage est lié au choix de la durée de l'intervalle de temps pendant lequel le signal détecté est analysé en vue de la classifier. Une première approche présentée en Section 5.1 consiste à choisir un intervalle à durée fixe. Cette durée est choisie comme étant la longueur maximale des sons à détecter. Cette solution a le désavantage d'envoyer au système de classification le son à identifier plus une partie du bruit environnemental plus ou moins importante.

Dans la section 5.2 une solution de détection de la fin du signal pour améliorer les performances du système de classification sera proposée.

Les deux approches de couplage entre les deux systèmes sont présentées dans les sections 5.1 et 5.2 avec leurs performances et la méthodologie d'évaluation du couplage détection-classification est présentée dans la section 5.3.1. La méthodologie d'évaluation de l'ensemble du système est présentée dans la section 5.4. Le chapitre se termine par la proposition d'une méthode de détection de sons clés en une seule passe (section 5.5).

5.1 Première approche : durée fixe du signal à partir de la détection

La première approche du couplage entre la détection et la classification réside dans l'utilisation d'une durée fixe du signal résultant de la détection.

La détection des événements sonores réalise l'étiquetage du signal. Après chaque détection on extrait une durée fixe de signal. Tous ces signaux de durée fixe sont envoyés au système de classification qui les identifie parmi les sept classes de sons. La durée est fixée à 7 secondes, durée maximale des sons à détecter.

Dans le cas des sons courts extraits avec une durée fixe, la durée du bruit environnemental qui suit le signal est plus importante que celle du signal lui-même, ce qui est un problème pour la classification (voir la figure 5.2).

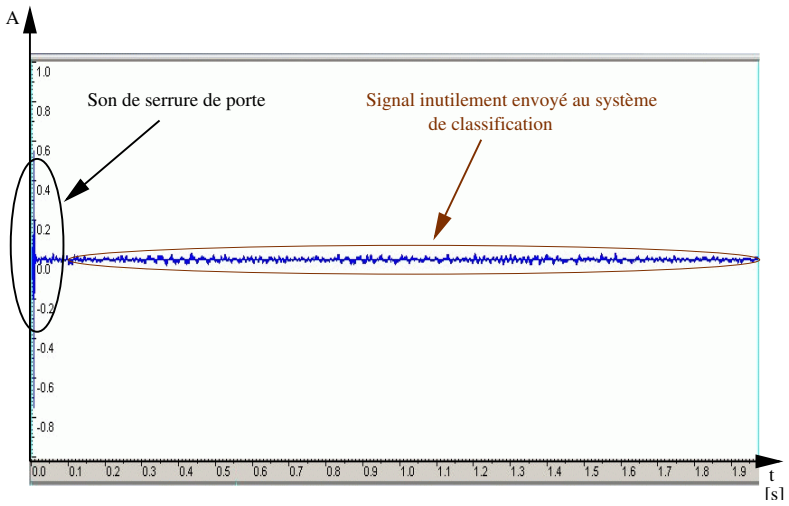


Fig. 5.2: Son d'une serrure de porte extrait avec une longueur fixe par le système de détection

5.2 Détection de la fin du signal avec l'algorithme fondé sur la transformée en ondelettes

L'utilisation d'une durée fixe pour les signaux détectés n'est pas adaptée à la classification des sons, ainsi une détection de la fin du signal est proposée dans ce paragraphe.

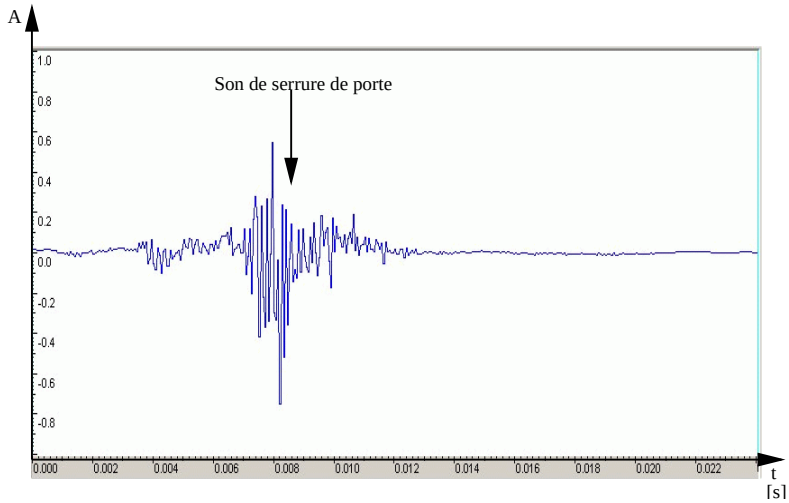


Fig. 5.3: Son d'une serrure de porte délimité par la détection du début et de la fin du signal

Pour détecter la fin du signal, l'algorithme de détection fondé sur la transformée en ondelettes est appliqué sur le signal inversé dans le temps [Press et al., 2002]. Le signal avec durée fixe est inversé dans le temps et l'algorithme de détection est appliqué encore une fois [Truchetet, 1998]. Après la détection de la fin du signal, celui-ci est extrait en utilisant le début indiqué par la première détection et la fin indiquée par la deuxième détection [Mallat, 2000]. Les signaux obtenus constituent l'entrée du système de classification. Dans la figure 5.3 nous avons le même son que précédemment, mais obtenu avec le système de détection de la fin du signal.

La détection de la fin du signal a comme hypothèse le fait que le signal utile a une durée plus courte que celle fixée ; il s'agit alors d'un type de signal qui ne fait pas partie des classes de sons identifiable par la classification donc une fausse alarme. La détection de la fin du signal a été validé sur la même base de test ;les mêmes performances que pour le début du signal ont été obtenues.

5.3 Évaluation des deux approches de couplage proposés

5.3.1 Méthodologie d'évaluation du couplage détection - classification

Pour évaluer le couplage du système de détection avec celui de classification, un corpus spécifique a été réalisé (voir section 2.4 de la page 46). Nous avons deux corpus de test : un premier avec le RSB variant de 0 à 10 dB (COUPLAGE1) uniformément réparti et un deuxième avec le RSB variant de 0 à 40 dB (COUPLAGE2) en mode aléatoire répartis d'après une loi de répartition qui correspond aux mésures effectués dans l'appartement.

La détection est effectuée avec l'algorithme fondé sur la transformée en ondelettes (le meilleur algorithme considéré). Le seuil de détection a été fixé à une valeur de 0.0007 consi-

dérée optimale après l'évaluation de l'algorithme sur la base de test de détection (voir section 3.3.3 de la page 80).

La classification des sons est effectuée avec le même système à base de GMM (4 gaussiennes) utilisant les MFCC combinés avec le nombre de passages par zéro, le roll-off point, le centroïde et l'énergie comme paramètres acoustiques. Le taux moyen d'erreur de classification est systématiquement calculé pour la classification des sons issus du système de détection automatique mais aussi pour les sons résultant d'une détection idéale (l'étiquetage des fichiers de la base de couplage est alors utilisé) ceci afin de voir l'influence des erreurs de détection sur le système complet. L'apprentissage des modèles GMM est effectué sur les sons purs de la base de classification avec le protocole «leave-one-out» déjà décrit (section 4.2.4.2 de la page 114).

5.3.2 Performances du couplage

Dans le tableau 5.1 sont résumées les performances de classification obtenues pour une détection idéale (étiquettage manuel du début et de la fin du signal), une détection réelle (ettiquettage donné par l'algorithme de détection) avec une durée fixe du signal et pour une détection avec durée estimée du signal par l'algorithme de détection du début et de la fin.

	Erreur moyenne de classification [%] RSB∈[10,20] dB	Erreur moyenne de classification [%] RSB∈[0,40] dB
Détection idéale TDM=0, TFA=0 durées réelles des signaux	21.5	22.7
Détection réelle TDM=1%, TFA=1% RSB∈(10,20) dB TDM=5%, TFA=2.6% RSB∈(0,40) dB durées fixes des signaux	67.8	69
Détection réelle TDM=1%, TFA=1% RSB∈(10,20) dB TDM=5%, TFA=2.6% RSB∈(0,40) dB durées estimées des signaux	**27.7**	**25.5**

Tab. 5.1: Les performances des deux approches de couplage entre la détection et la classification

Les performances de la classification couplée avec un système de détection idéale confirment les résultats de classification en présence du bruit donnés dans le chapitre 4.

Les performances de la classification des sons détectés et extraits avec une durée fixe de 7 secondes sont inacceptables. La grande différence (46.3%) au niveau de l'erreur de classification entre les deux premières lignes du tableau montre que ce couplage n'est pas adapté au système de classification.

La détection de la fin du signal améliore significativement (de 40.1%) les performances de la classification pour une détection réelle. La différence entre ce couplage et le système de détection idéal reflète l'influence des fausses alarmes et des détections manquées sur l'algorithme de classification.

5.4 Évaluation du système global détection - classification pour la détection des situations de détresse

5.4.1 Méthodologie

L'évaluation de l'ensemble du système est difficile parce que les deux sous-systèmes (détection et classification) ont des caractéristiques différentes. Les erreurs possibles du système de détection d'événements sonores sont les fausses alarmes (la détection d'événements inexistants) et les détections manquées (la non-détection de vrais événements sonores). Les taux des détections manquées et des fausses alarmes sont définis dans la section 3.2.1.4 (page 70). Le système de classification des sons peut produire des erreurs de classification qui n'ont pas toutes des conséquences négatives pour la détection des situations de détresse.

Les 7 classes de sons à identifier peuvent être divisées en : classes de sons qui génèrent une alarme (A) et classes de sons qui ne génèrent pas d'alarme ($\overline{A}$). Nous pouvons définir pour l'ensemble du système une *détection manquée globale* comme étant la non détection d'un son appartenant à une classe de type A et une *fausse alarme globale* comme la détection d'un son de type A alors qu'aucune alarme n'était présente en entrée.

Ces définitions nous amènent à la matrice de confusion de l'ensemble du système présentée dans le tableau 5.2.

	Classe en entrée / Classe reconnue	$\overline{A}$				A		
		C1	C3	C4	C9	C2	C5	C7
$\overline{A}$	Claquements de porte (C1)	✓	-	-	-	**DM**		
	Sonneries de téléphone (C3)	-	✓	-	-			
	Sons de pas (C4)	-	-	✓	-			
	Serrures de porte (C9)	-	-	-	✓			
A	Bris de verre (C2)	**FA**				✓	-	-
	Cris (C5)					-	✓	-
	Vaisselle (C7)					-	-	✓

Tab. 5.2: Matrice de confusion de l'ensemble du système (DM : Détection manquée, FA : Fausse Alarme, ✓ : Bonne classification, - : erreur de classification sans conséquences, A : alarme, $\overline{A}$: sans alarme)

Remarque : Les détections manquées de l'ensemble du système comprennent aussi les détections manquées de l'étage de détection.

5.4.2 Résultats

Sur la base des performances de la détection et de la classification, nous avons calculé le TDM global et le TFA global pour les deux bases de test. Une synthèse de ces résultats est présentée dans le tableau 5.3. Les résultats détaillés sont donnés dans l'annexe G (page 171).

Le Taux Global de Détections Manquées est de 3% pour les deux bases de test. La valeur de 3% de détection manquées est encore trop élevée pour une application de surveillance médicale, et doit être analysée en tenant compte du fait que la sortie du système d'analyse sonore

	RSB∈[10,20] dB	RSB∈[0,40] dB
TDM_{Global} [%]	3	3
TFA_{Global} [%]	12.3	12.7

Tab. 5.3: Performances globales du système de détection de situations de détresse

est fusionnée avec celles d'autres capteurs médicaux. Le Taux Global de Fausses Alarmes est approximativement de 12% pour les deux cas.

5.5 Vers une détection de sons-clés

Dans cette section nous présentons une première étude d'une approche différente d'identification de sons clés. Nous remplaçons l'analyse du signal en deux étapes (détection suivi de classification) par une analyse dans une seule étape : la recherche de sons clés.

Cette approche élimine l'étape de la détection des sons avant la classification par une recherche continue des sons à identifier dans le flux sonore. Elle est inspirée des techniques de «Word spotting» [Boite et al., 2000].

Seck propose pour la détection d'une classe de sons une modélisation de la classe cible et de la classe non-cible par un mélange de gaussiennes (GMM) [Seck, 2001]. L'information de la présence ou non de la classe cible est réalisée par seuillage du rapport des vraisemblances des trames à identifier en rapport avec les deux classes.

Pour le suivi d'une classe, le même auteur, décrit un système en deux étapes : une première étape qui consiste en détection des ruptures, suivi d'un seuillage des vraisemblances des trames avec les modèles de mélange de gaussiennes de chaque classe sonore à identifier. La vraisemblance est calculée sur une moyenne des vraisemblances de plusieurs trames. Les tests effectués sur un suivi de la parole dans une émission documentaire ont donné un taux d'égale erreur de 12% pour une fenêtre de lissage entre 2 et 5 secondes. Trois modèles sont utilisés : le modèle de la parole seule, le modèle de la parole sur fond musical et le modèle de la musique seule.

La modélisation des classes peut aussi être faite avec des modèles de Markov Caché (HMM) comme proposé dans [Gauvain et al., 1999] et [Meignier et al., 2000].

Dans l'algorithme proposé, le signal sonore est divisé en trames pour le calcul des vecteurs acoustiques. Un modèle statistique pour chaque classe sonore à identifier est obtenu par apprentissage. Une classe bruit environnemental est introduite. La vraisemblance d'appartenance de chaque vecteur acoustique en rapport avec chaque classe sonore est calculée. Un moyennage sur un nombre L de trames est effectué sur les vraisemblances en rapport avec chaque classe. La classe identifiée à chaque instant est obtenue par comparaison des vraisemblances de chaque trame en rapport avec toutes les classes sonores (voir figure 5.4).

5.5.1 Une première évaluation

Une première évaluation de l'algorithme de détection de sons-clés proposé a été effectuée. Le signal de test a une durée de 460 secondes et contient 10 sons à identifier pour diverses valeurs du RSB entre 10 et 20 dB, espacés par des moments de silence de durée aléatoire. Le bruit utilisé est le bruit de l'appartement HIS.

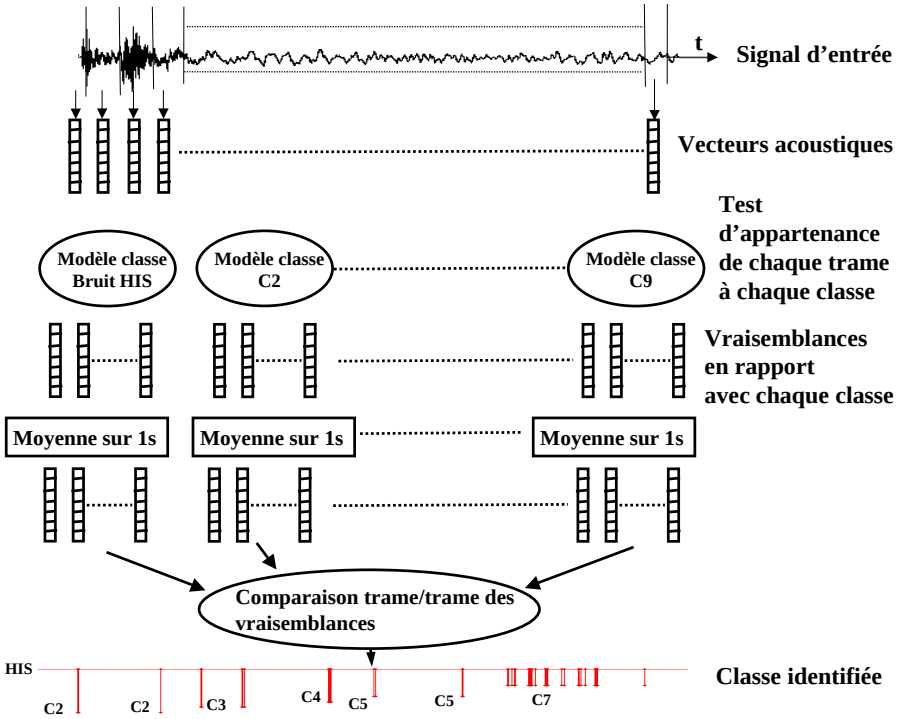


Fig. 5.4: L'organigramme du système de détection de sons clés

L'apprentissage des modèles est effectué sur les sons de la base de classification (voir section 2.3 de la page 44). Pour le bruit de l'appartement HIS, 300 secondes d'enregistrement sont utilisées.

Le vecteur acoustique est calculé sur des fenêtres de signal de 16ms (256 échantillons) avec un recouvrement de 50%. Les paramètres acoustiques utilisés sont les 16 MFCC en conjonction avec le nombre de passages par zéro, le Roll-off point, le Centroïde et l'énergie.

Le moyennage des vraisemblances est effectué sur L=12 trames ($\approx$ 1seconde) avec un déplacement d'une trame à chaque instant de temps (92% de chevauchement entre les fenêtres).

5.5.1.1 Résultats

Le résultat de détection présenté dans la figure 5.5 est obtenu en utilisant six classes de sons (c'est-à-dire les classes C2-C5, C7, et C9) et la classe du bruit de l'appartement HIS.

Ce premier résultat montre l'utilisabilité de cet algorithme. Nous pouvons observer que tous les 10 sons ont été détectés. Nous observons cependant 5 fausses alarmes et une mauvaise classification du son appartenant à la classe «Serrure de porte» (C9), qui est reconnu comme appartenant à la classe «Vaisselle» (C7).

Ces premières résultats semblent convaincants, mais seule une évaluation complète à partir des corpus d'évaluation COUPLAGE1 et COUPLAGE2 permettra de conclure sur les performances de la méthode de détection de sons-clés.

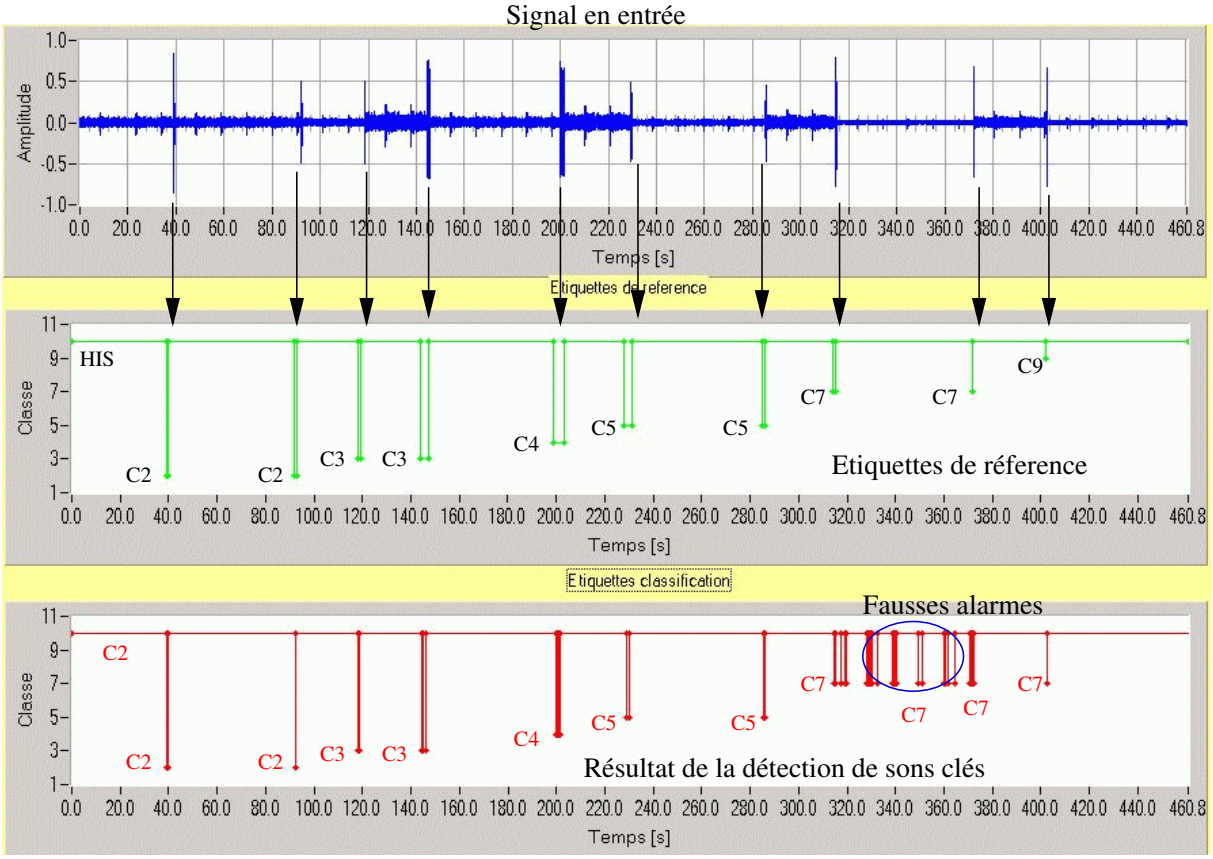


Fig. 5.5: Résultats de détection de sons-clés sur un signal de 400s avec 10 sons à identifier

Bibliographie

[Boite et al., 2000] Boite, R., Bourlard, H., Dutoit, T., Hang, J., and Leich, H. (2000). *Traitement de la parole*. ISBN 2-88074-388-5. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Dufaux, 2001] Dufaux, A. (2001). *Detection and recognition of Impulsive Sounds Signals*. PhD thesis, Faculté des sciences de l'Université de Neuchâtel, Suisse.

[Gauvain et al., 1999] Gauvain, J., Lamel, L., and Adda, G. (1999). Audio partitionning and transcription for broadcast data indexation. *First European Workshop on Content-Based Multimedia Indexing CBMI'99*, pages 67–73, Toulouse, France.

[Mallat, 2000] Mallat, S. (2000). *Une exploration des signaux en ondelettes*. ISBN 2-7302-0733-3. Les Editions de l'Ecole Polytechnique, Paris.

[Meignier et al., 2000] Meignier, S., Bonastre, J., Fredouille, C., and Merlin, T. (2000). Evolutive HMM for multi-speaker tracking system. *International Conference on Audio, Speech and Signal Processing ICASSP '2000*, pages 543–547, Istanbul, République Turque.

[Press et al., 2002] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (2002). *Numerical Recipes in C ; The Art of scientific Computing ;The second Edition*. ISBN 0-521-43108-5. Cambridge University Press.

[Seck, 2001] Seck, M. (2001). *Détection de ruptures et suivi de classe de sons pour l'indexation sonore*. PhD thesis, Université de Rennes I.

[Truchetet, 1998] Truchetet, F. (1998). *Ondelettes pour le signal numérique*. ISBN 2-86601-672-6. Hermes.

Chapitre 6

Conclusions et perspectives

Depuis quelques années se développe le concept général d'espace perceptif ou salle intelligente qui répond de diverses façons aux besoins, demandes, attentes des acteurs humains. Les espaces perceptifs traitent des signaux de parole, des signaux vidéo, des données de l'environnement, de la localisation des personnes, du suivi et de la reconnaissance des gestes, etc.

Ce travail de thèse se situe à la frontière entre les espaces perceptifs et la télémédecine qui a récemment évolué vers la téléchirurgie, la télésurveillance des malades, le télédiagnostic, etc. La télésurveillance est notamment utilisée pour suivre l'évolution de personnes à risques (maladies chroniques ou personnes exposées à des situations critiques). Cela peut être à domicile (personnes âgées) ou dans un environnement professionnel dangereux.

L'analyse et l'extraction des informations du son est un aspect important des espaces perceptifs pour la télésurveillance médicale. Dans ce contexte, cette thèse analyse et propose des solutions aux problématiques qui sont spécifiques au traitement du son dans les espaces perceptifs plus particulièrement pour la télésurveillance médicale.

Une problématique importante des espaces perceptifs, qui a été abordée, est la qualité et la quantité des signaux traités. Tous les algorithmes étudiés et proposés dans ce manuscrit ont été évalués et validés sur des signaux avec un rapport signal sur bruit variant dans une large gamme (de 0 dB à 60 dB). La grande quantité des signaux à traiter en temps réel (150 Ko/s) a imposé l'utilisation d'algorithmes rapides mais ayant néanmoins des performances acceptables pour une application de télésurveillance médicale.

Une autre problématique qui a été traitée dans ce manuscrit est celle de la méthodologie d'évaluation des performances de chaque partie du système et du système global. Une adaptation des méthodologies existantes dans la reconnaissance de la parole/du locuteur aux problématiques des espaces perceptifs a été proposée. Pour chaque partie du système proposé une étude de l'existant a été effectuée.

Une des contributions de cette thèse est la proposition de trois algorithmes de détection des événements sonores dans le bruit. L'adaptation des méthodes de reconnaissance de la parole/du locuteur à la classification des sons de la vie courante constitue une autre problématique abordée. L'étude s'est concentrée sur la recherche de paramètres acoustiques mieux adaptés aux sons de la vie courante.

Le système global proposé est constitué de deux parties : la première est la détection des événements sonores et la deuxième est la classification des sons. Une raison de la division du

système en deux parties est le temps de calcul important des algorithmes de classification qui ne peuvent pas être appliqués sur un flux audio continu multivoies.

Les principales contributions du travail de thèse sont décrites ci-dessous.

La création d'un corpus des sons de la vie courante. Dans le chapitre 2 sont présentées les problématiques de l'enregistrement d'un corpus de sons de la vie courante. L'inexistence d'un corpus des sons de la vie courante nous a amené à enregistrer un tel corpus (durée = 1 heure et 35 minutes). L'enregistrement a été effectué en utilisant la méthodologie des corpus de parole. Les fichiers d'étiquettes sont dans le standard SAM adapté à notre tâche. Ce corpus a été le point de départ de la réalisation de trois autres bases de test pour l'évaluation du système de détection, du système de classification et du couplage entre les deux.

La détection des événements sonores dans le bruit. La détection des événements sonores est abordée dans le chapitre 3. L'étude commence par une évaluation des algorithmes issus de l'état de l'art sur la base de test réalisée. Les performances de ces algorithmes étant inacceptables pour notre application, trois nouveaux algorithmes de détection sont proposés. Le premier étant basé sur la fonction d'intercorrélation des deux fenêtres successives du signal a des performances très bonnes pour le bruit réel mais inacceptables pour le bruit blanc. Le deuxième algorithme proposé se fondant sur la prédiction de l'énergie du signal par des fonctions SPLINE présente de performances moyennes mais il est très rapide. Le dernier algorithme proposé se fonde sur la transformée en ondelettes du signal. Cet algorithme présente de très bonnes performances pour le bruit réel (un taux d'égal erreur de 5.6% pour un RSB=0 dB et de 0% pour les autres cas) et pour le bruit blanc. Ce troisième algorithme proposé semble mieux adapté à notre application où les sons sont majoritairement impulsionnels. Le seuil de cet algorithme est adaptatif en fonction du niveau du bruit. Les algorithmes ne sont pas sensibles à la position du son dans la fenêtre d'analyse (qui a une taille réduite $\approx$ 128 ms) ;cette affirmation a été validé par l'évaluation des algorithmes de détection sur les bases de test COUPLAGE1, COUPLAGE2.

Dans le cas concrète de l'application de télésurveillance médicale à domicile, les algorithmes de détection ont été conçus pour pouvoir détecter un son même dans la présence d'un bruit extérieur modéré (comme le bruit de la rue). En ce qui concerne la présence d'un poste de télévison ou de radio utilisation d'un canal supplémentaire d'acquisition est prévue ; ce canal servira à une soustraction du signal parasite des signaux des voies de surveillance.

La classification des sons - les paramètres acoustiques. Le système de classification des sons est présenté dans le chapitre 4. L'étude des méthodes existant dans le domaine de la reconnaissance des formes et plus particulièrement dans celui de la reconnaissance de la parole/du locuteur nous a amené à choisir un algorithme à base de mélange de gaussiennes (GMM). Comme la paramétrisation du signal est très importante pour une bonne classification des sons, l'étude des paramètres acoustiques adaptés aux sons de la vie courante a constitué l'axe principal de recherche. La méthodologie consiste à évaluer par des méthodes statistiques la pertinence des paramètres utilisés habituellement pour la détection de signaux musicaux (le nombre de passages par zéro, le Roll-off point et le Centroïde). Étant donné que la transformée en ondelettes est mieux adaptée à l'analyse de signaux impulsionnels, la possibilité d'extraire des paramètres acoustiques à partir de cette transformée a été aussi étudiée.

La deuxième étape de la méthodologie est l'analyse du comportement du système de classification en présence du bruit. Les deux approches envisagées sont l'utilisation de paramètres acoustiques «résistants au bruit» et le pré-traitement des signaux. Un pré-traitement fondé sur le seuillage adapté des coefficients de la transformée en ondelettes a été proposé et validé. L'amélioration apportée par ce pré-traitement est de 8% sur le taux d'erreurs de classification.

Le couplage des deux systèmes. Le couplage de la détection des évènements sonores avec la classification des sons de la vie courante, présenté dans le chapitre 5 est critique pour la phase de classification. La précision de détection du signal qui sera envoyé au système de classification est un paramètre important pour la performance de l'ensemble. Le problème est donc de détecter aussi la fin du signal. Une solution est proposée en appliquant l'algorithme de détection sur le signal inversé dans le temps. Avec cette solution de couplage, les performances de classification du système complet restent proches de celles du système utilisant une détection parfaite manuelle (6% de perte seulement).

L'évaluation de l'ensemble du système. La méthodologie d'évaluation du système d'extraction des informations sonores pour une application de télésurveillance médicale est aussi présentée dans le chapitre 5. La méthode proposée tient compte des influences des erreurs possibles du système sur l'application de télésurveillance médicale. Avec cette méthode d'évaluation, nous obtenons un taux global de détection manquées de 3% et un taux global de fausses alarmes de 12%. Ces performances doivent être analysé tenant compte du fait que ce système est utilisé en conjonction avec d'autres capteurs par l'intermédiaire d'une fusion de données, qui ne faisait pas l'objet de ce travail de thèse. Géneralement le système de télésurveillance médicale a été conçu un grand degré de redondance. L'amélioration des performances globales du système dépend de l'algorithme de fusion de donnée.

Perspectives

La détection des événements sonores, la première étape du système, dispose d'un algorithme fondé sur la transformée en ondelettes qui peut être amélioré en passant de l'analyse faite par l'énergie de chaque coefficient de la transformée en ondelettes à une analyse statistique de l'évolution de ces paramètres. Pour l'instant, l'algorithme proposé n'utilise que la décomposition en fréquence de la transformée en réalisant une moyenne de l'évolution temporelle (par le calcul de l'énergie). La prise en compte de l'évolution temporelle de la transformée en ondelettes (variance des coefficients, paramètres dérivée des coefficients, etc.) pourrait permettre une meilleure précision de détection.

L'étape de classification réalisée avec un système à base de GMM sur des signaux bruités peut être aussi améliorée, soit par l'amélioration du pré-traitement, soit par un apprentissage sur une base de test bruitée. L'apprentissage des modèles GMM sur les signaux purs et bruités est actuellement utilisé dans le domaine de la reconnaissance de la parole/du locuteur avec de bonnes performances. Le pré-traitement proposé est fondé sur la transformée en ondelettes et il pourra être amélioré par un seuillage adaptatif des coefficients de la transformée en ondelettes (actuellement le seuil est fixé définitivement au départ sur le début du signal).

En opposition, dans la section 5.5 un système à une seule étape est proposé. Ce système évite la détection des événements sonores en procédant à la classification de chaque trame de

signal parmi les classes de sons de la vie courante et celle du bruit environnemental. Ce système doit être évalué et validé sur la base de test existante. La possibilité d'intégrer un tel algorithme dans un environnement temps réel reste à étudier. Par ailleurs, une étude approfondie de l'état de l'art dans les domaines de détection de mots clés (Word Spotting) et de locuteur (Speaker Tracking) devrait nous permettre d'adapter ses techniques à la détection des sons comme dans le cas de la classification de sons présentée dans ce manuscrit.

Ces recherches montrent que l'utilisation des informations extraites du son peut contribuer à la robustesse et à l'efficacité d'un système de télésurveillance médicale. La télésurveillance des personnes âgées à domicile réduit le temps d'intervention en cas d'accidents domestiques ou de malaises. La présence du son dans un tel système offre la possibilité d'identifier des situations de détresse comme une chute pour une personne qui a oublié son capteur de chute. L'identification des sons de la vie courante peut aussi faciliter la détection des pathologies de diverses maladies. Une première étude de fusion de données pour l'identification de certaines pathologies est en cours (pour chaque pathologie à identifier ont été établis des scénarios à reconnaître par le système).

Le système présenté peut être facilement adapté à la télésurveillance dans les systèmes d'aide aux personnes à mobilité réduite qui represente une application de télésurveillance médicale.

Les algorithmes développés dans cette thèse pourraient être aussi appliqués à des systèmes de contrôle d'accès de bâtiments pour des applications professionnelles (accès à des bureaux, etc.) ou destinées à des particuliers (habitations privées). En effet, la présence d'une détection et identification de son peut éliminer les fausses alarmes des capteurs classiques de surveillance (par exemple : l'identification d'un son suspect couplé avec la détection d'un mouvement par un système classique).

Une autre application possible est la domotique, en proposant un espace perceptif à domicile, répondant aux besoins des habitants. Le son pourrait permettre une interaction plus conviviale entre les habitants et un système contrôlant divers appareils domotiques.

Le système présenté peut être utilisé dans les systèmes d'indexation des archives sonores ou vidéo ou dans des systèmes domotiques comme «objet communicant» d'analyse sonore.

ANNEXES

Chapitre A

Rappels mathématiques

A.1 Méthode de calcul des coefficients LPC basée sur l'autocorrélation

Le calcul des coefficients LPC nécessite le calcul de l'autocorrélation biaisée d'une suite $s(n)$ de N échantillons :

- ajout de N zéros à la fin du signal pour supprimer la périodisation du signal en calculant sur une fenêtre de taille double de la taille réelle :

$$x(n) = \begin{cases} s(n) & si \quad 0 < n < N \\ 0 & si \quad N \leq n < 2N \end{cases}$$

- utilisation de la TFR(Transformée de Fourier Rapide, en anglais FFT) pour le calcul de l'autocorrélation :

$$\Gamma_x = \frac{1}{2N-1}.TFR^{-1}[|X(\nu)|^2]$$

L'autocorrélation biaisée du signal étant calculée, on utilise l'algorithme classique dit de Durbin-Levinson pour obtenir les M coefficients LPC $a(m)$ [CALLIOPE, 1989].

L'algorithme est itératif, il se décompose de la façon suivante en désignant l'étape de calcul par i et en utilisant les variables internes k, p et $a_i(m)$:

- étape d'initialisation ($i = 0$)

$$\begin{cases} p = \Gamma_x(0) \\ a_0(0) = 1 \end{cases}$$

- M itérations, $i$variant de 1 à M

$$\begin{cases} k = \frac{1}{p}\left[\Gamma_x(i) - \sum_{j=1}^{j<i} \Gamma_x(i-j).a_{i-1}(j)\right] \\ p = p.(1-k^2) \\ a_i(i) = k \\ a_i(j) = a_{i-1}(j) - k.a_{i-1}(i-j) \qquad 1 \leq j \leq i-1 \end{cases}$$

– résultats du calcul

$$\begin{cases} a(0) = 1 \\ a(1) = a_M(1) \\ \quad \vdots \\ a(m) = a_M(m) \\ \quad \vdots \\ a(M) = a_M(M) \end{cases}$$

A.2 Détermination des dérivées des coefficients (Δ, $\Delta\Delta$)

A.2.1 Dérivée première (Δ).

Comme la fonction de variation des paramètres acoustiques est connue seulement en des instants précis, le calcul de la dérivée première doit être fait par approximation [Press et al., 2002]. Ceci nécessite de connaître au moins 2 valeurs de la fonction autour du point concerné, les calculs se simplifient en utilisent 5 valeurs régulièrement espacées. Les 2 valeurs précédant la valeur courante c_k sont c_{k-2} et c_{k-1}, les 2 valeurs suivantes c_{k+1} et c_{k+2}. La formule d'approximation de la dérivée première (A.1) s'obtient simplement à partir de la décomposition en série de Taylor de la fonction.

$$\Delta c_k = \frac{-(c_{k+2} - c_{k-2}) + 8.(c_{k+1} - c_{k-1})}{12} \tag{A.1}$$

A.2.2 Dérivée seconde ($\Delta\Delta$).

La formule de calcul de la dérivée seconde à partir de 5 valeurs régulièrement espacées est obtenue avec le même développement en série Taylor que celui utilisé pour la dérivée première.

$$\Delta\Delta c_k = \frac{-(c_{k-2} - 16.c_{k-1} + 30.c_k - 16.c_{k+1} + c_{k+2})}{12} \tag{A.2}$$

Chapitre B

Divers projets de télémédecine

Parmi les principaux projets de télémédecine internationaux, on citera notamment :

- ✓ **Columbia** qui s'occupe de la détection de l'aggravation des crises d'asthme pour des asthmes instables traités au domicile. Le projet est mis en oeuvre par l'hôpital de la capitale de l'état de Caroline de Sud (États-Unis). Les données respiratoires sont recueillies au domicile via un spiromètre raccordé à un PC qui les transmet directement sur le réseau (Web). A l'hôpital, les données sont enregistrées dans le dossier médical du patient. Le médecin traitant détermine pour chaque patient les seuils d'alarme. C'est une application où on transmet un seul type de données médicales de type *télésurveillance*.
- ✓ **VBCH Telemedicine project** Ce projet est proposé par l'Université de Iowa (USA) et l'hôpital «Van Buren» et a comme but la communication vidéo et sonore entre les patients et le personnel médical [VBCH Project, 2002]. C'est un système de *transmission des images et des sons* qui n'a aucune spécificité technique particulière.
- ✓ **UIUC Dysphagia Telemedicine project** de l'Université Illinois a comme but le développement d'outils pour permettre aux experts médicaux de réaliser en temps réel, avec une grande résolution, télécommandé, la vidéo-fluoroscopie des malades d'oral/pharyngéal dysphagie [UIUC Project, 2002]. Le projet entre dans la thématique de la *transmission d'un flux vidéo important* et de la *télématique*.
- ✓ **UWGSP9 Telemedicine project** Le système UWGSP9 développé par le laboratoire ICSL[1] à l'Université de Washington, donne aux experts médicaux et aux spécialistes situés à des endroits différents, la possibilité de consulter les patients à distance à travers un mélange de vidéo, audio et images externes [Lee et al., 1994]. Les chercheurs du laboratoire ICSL ont conçu une station de travail basée sur un DSP de type TMS 320C80. La spécificité technique reste au niveau de l'acquisition, la compression et la transmission de la vidéo et du son. Ce projet fait partie des applications de type *télédiagnostic*.
- ✓ **LITMED II** Ce projet est proposé par Kaunas University of Technology et Informations Logik AB avec d'autres partenaires [LITMED Project, 2002]. C'est un projet qui réalise la liaison entre les personnels médicaux de Lituanie, la Suède et d'autres pays Baltes. Le but principal est de faciliter les échanges en temps réel entre les spécialistes de ces pays. C'est une application de type *télé-éducation*.

[1] Image Computing System Lab

- ✓ **DROMEAS** est un projet européen entre des entreprises des pays suivants : Angleterre, Grèce, Chypre, Belgique, République Tchèque et Israël. Son but est la réalisation d'une station portable qui mesure des grandeurs médicales d'un athlète en mouvement et les transmet à une station de base [Vassiliadis, 2003]. C'est un système de *télésurveillance* qui implique des capteurs physiologiques, des algorithmes d'analyse en temps réel et une transmission par ondes radio.
- ✓ **Système automatique de distribution des pilules** pour les personnes âgés proposé par l'Université de Tromsö, Norvège [Olsen et al., 2003]. Cet appareil distribue au malade les pilules à prendre, en fonction de la programmation qui se fait par réseau Internet. Il transmet au centre médical un rapport horodaté des médicaments distribués. C'est une application de *télésurveillance* automatique de la distribution des médicaments réalisé en Norvège.
- ✓ **Philips - les outils de télésurveillance** Philips produit des appareils de mesure de la tension artérielle et du pouls, du poids, etc., qui transmettent les mesures vers une station centrale qui peut se trouver dans un hôpital. Philips produit des appareils de mesures nécessaires pour les applications de type télésurveillance.
- ✓ **Projet TIISSAD - Technologies de l'Information Intégrées aux Services des Soins à Domicile** est une collaboration entre IUP Paris XIII, LORIA Altir, INSERM ERM 107 Lyon, INSERM USS8 Toulouse et TIMC-IMAG. Le but du projet a été l'utilisation des nouvelles technologies de l'information pour la télémedecine. Parmi les résultats du projet : système de télésurveillance ECG, diagnostic pour les maladies des reins et surveillance de mouvements pour les malades souffrant de la maladie d'Alzheimer. C'est une application de télésurveillance du rythme cardiaque, de la position et de fusion de données [Projet TIISSAD, 2001].

Chapitre C

Système matériel pour l'analyse multivoies du son

C.1 Appartement HIS

L'appartement utilisé pour notre étude est situé dans les locaux du laboratoire TIMC. C'est un appartement de type 2 pièces, de surface totale 30 m^2. Il dispose d'un local technique séparé de l'appartement par une vitre sans tain. Dans ce local sont installés les ordinateurs du système de surveillance ainsi que les instruments d'acquisition. La vitre permet de suivre les activités dans l'appartement. L'appartement est équipé de différents capteurs :

- sonores : 5 microphones omni-directionnels sont disposés dans chacune des pièces (couloir, douche, toilettes, cuisine et séjour) ; le microphone du séjour capte aussi les sons de la chambre (conformément au schéma de l'appartement de la figure 1.2 - page 28)
- infra-rouge de position : 5 capteurs volumétriques dans les pièces : couloir, douche, cuisine, séjour et chambre
- contacts d'ouverture/fermeture sur la porte d'entrée et la porte des toilettes
- pèse personne
- actymètre indiquant la position de la personne
- oxymètre
- tensiomètre

L'architecture envisagée du système de surveillance médicale est présenté dans la figure C.1. Le système de télésurveillance complet est constitué de trois PC reliés entre eux par un bus CAN et une connexion Ethernet. Le PC «maître» a en charge l'acquisition des signaux provenant des capteurs physiologiques, des capteurs de position et du «*capteur sonore intelligent*». Le dialogue entre le PC «maître» et les capteurs passe par un bus CAN. Le PC «maître» réalise la fusion de données issues de tous les capteurs et prend la décision d'appeler les urgences dans le cas d'une possible situation de détresse.

Le capteur sonore intelligent est composé de deux PC : l'un a pour tâche l'acquisition en continu des 5 canaux sonores, la détection des événements sonores, la classification des sons de la vie courante et la communication par réseau Ethernet avec le deuxième PC qui, quant à lui, réalise la reconnaissance des expressions de détresse. *Le capteur sonore intelligent* transmet

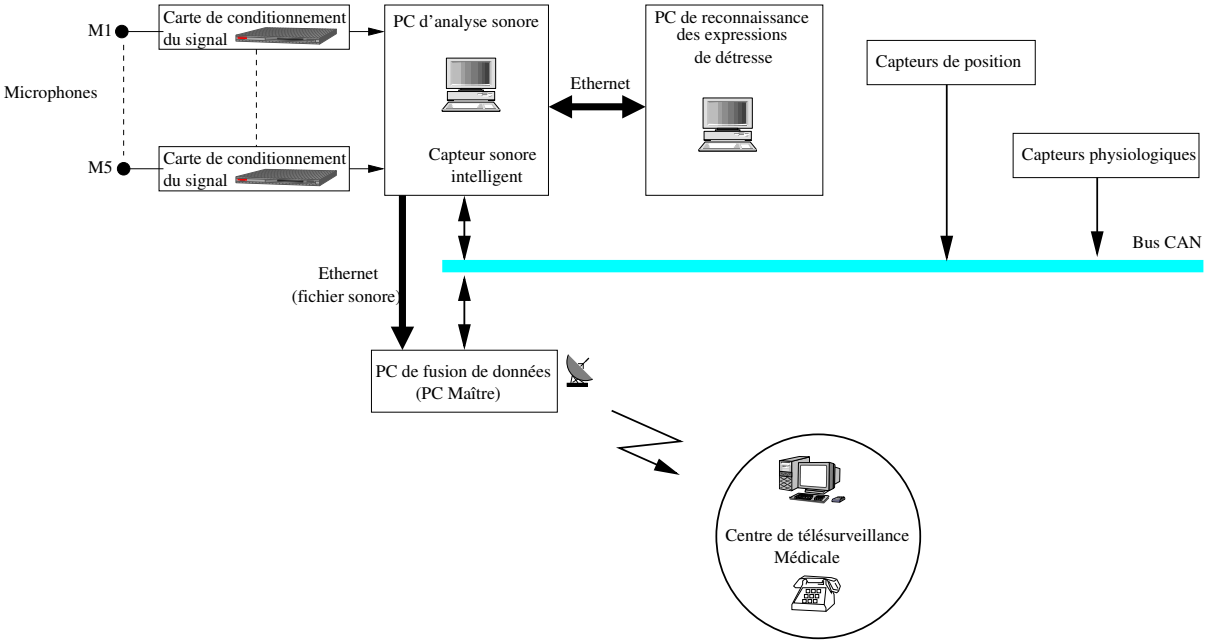


Fig. C.1: Architecture matérielle du système de surveillance médicale

au PC «maître» les informations extraites de l'analyse du son aussi bien en ce qui concerne la reconnaissance des situations de détresse que la classification des sons de la vie courante. Le fichier sonore considéré comme «suspect» est transmis pour expertise par une liaison Ethernet. La durée du fichier sonore étant de 10 secondes, le signal n'est pas compressé.

Le schéma envisagé pour le système d'analyse sonore est présenté dans la figure C.2. Chacun des 5 canaux sonores est associé à un processus de détection qui fonctionne en permanence. Dans le cas de détection d'un événement sonore, le signal est extrait et envoyé au système de segmentation sons de la vie courante / parole. Après segmentation, soit le système de classification des sons de la vie courante, soit le système de reconnaissance des expression de détresse est activé. Un son de la vie courante est classé parmi les 7 classes. Si l'événement sonore est classé dans une classe de sons considéré comme «à risque», ou si une expression de détresse est reconnue, un message d'alarme est envoyé vers le PC «maître».

Le système de détection d'événements sonores est actuellement implémenté dans l'habitat HIS. La classification des sons de la vie courante et la reconnaissance des expressions de détresse sont en cours d'implémentation. La détection et la classification des sons de la vie courante font partie des objectifs de ce travail.

Au niveau matériel, le PC d'acquisition sonore est équipé d'une carte d'acquisition multivoies (National Instruments) et d'une carte PCI-CAN pour la connexion au bus CAN. Chaque microphone est interfacé à la carte multivoies par une carte de conditionnement du signal (amplification avec un amplificateur de mesure et filtrage anti-repliement).

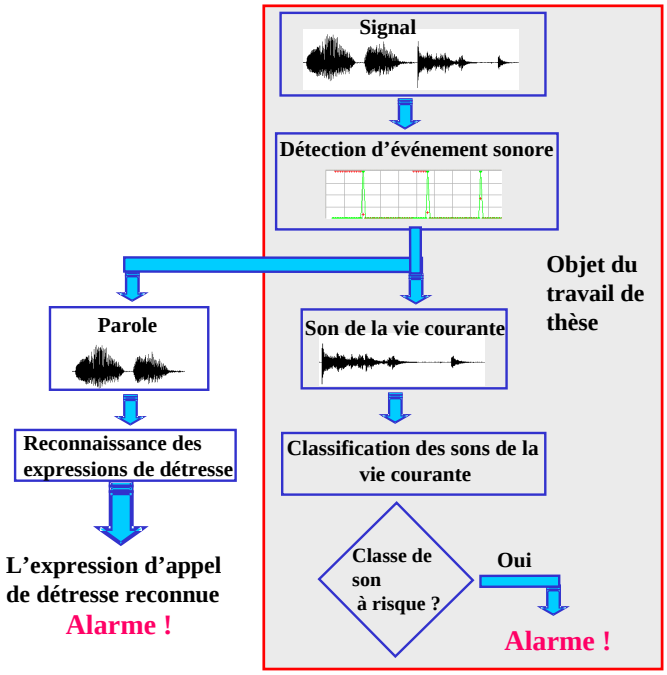


Fig. C.2: Architecture du système d'analyse sonore

C.2 Carte d'acquisition

La carte d'acquisition utilisée est de type PCI 6034E de National Instrument (voir figure C.3) [N.I., 1999f]. Cette carte dispose de 8 entrées différentielles ou 16 entrées non différentielles avec une fréquence maximale d'échantillonnage de 200 kéchantillons/s ce qui nous permet d'utiliser une fréquence d'échantillonnage maximale de 25 kHz/canal (pour 8 canaux d'acquisition sonore) [N.I., 1999a]. L'acquisition des canaux sonores est faite par multiplexage temporelle. La carte dispose aussi de 8 entrées/sorties numériques non utilisées par cette application.

Fig. C.3: Carte d'acquisition PCI 6034E

Nous utilisons les 8 entrées analogiques différentielles de la carte avec une fréquence d'échan-

tillonnage de 16KHz/canal qui est la fréquence habituellement utilisée dans la reconnaissance de la parole.

La carte est programmée avec LabWindows/CVI qui est un atelier de développement ANSI C incluant des bibliothèques graphiques, de traitement du signal, etc. L'acquisition, en continu, sur les 8 canaux, à une fréquence d'échantillonnage de 16KHz/canal avec sauvegarde sur le disque dur s'est avérée impossible en temps réel avec les fonctions de haut-niveau fournies par National Instrument. L'acquisition utilise donc les fonctions de bas-niveau de la carte pour accéder directement aux valeurs de sortie du convertisseur analogique-numérique. Les valeurs sont sauvegardées dans la mémoire du PC par un processus de type DMA avec transfert par bloc de 2048 échantillons en utilisant une technique de tampon double. Dans le cas du lancement d'un algorithme de détection, les calculs de détection sont réalisés pour chaque trame avant l'arrivée de la suivante. Lors d'un enregistrement de corpus, le logiciel sauvegarde les échantillons, en continu sur le disque dur du PC.

C.3 Microphones

Nous avons choisi d'utiliser des microphones à condensateur parce qu'ils sont de dimensions réduites et omni-directionnels. Le signal électrique de sortie de ce type de microphone a une amplitude réduite qui ne couvre pas la plus petite gamme d'entrée de la carte d'acquisition (-100 mV, 100 mV). Nous avons donc réalisé pour chaque microphone une carte de conditionnement du signal [PROTEL, 1999]. La carte de conditionnement du signal réalisée amplifie le signal ; un filtre passe-bas a été introduit parce que la carte d'acquisition n'est pas équipée d'un filtre anti-repliement. La carte de conditionnement utilise un amplificateur opérationnel de type SSM2017 (Analog Device) et un filtre anti-repliement à capacités commutées de type MF6CN-50 (filtre Butterworth d'ordre 4).

Les cartes de conditionnement du signal ont été installées dans le faux plafond de l'appartement et le signal est transmis par câble blindé.

C.4 Liaison par bus CAN

Le PC d'analyse sonore et le PC réalisant la fusion de données communiquent à travers un bus Controller Area Network (CAN) [CAN Bus site, 2003]. Le bus CAN, est un bus série défini par la norme ISO 11898. Tous les dispositifs reliés à un bus CAN peuvent communiquer entre eux par l'intermédiaire d'une paire torsadée. En cas de collision, le bus CAN donne une réponse déterministe au contraire d'un bus Ethernet. Chaque noeud d'un bus CAN se voit attribuer un niveau de priorité ; dans le cas d'une collision le noeud ayant le niveau de priorité le plus élevé **continue** à transmettre, l'autre interrompt instantanément son émission. Le bus CAN est facile à implémenter et bien protégé contre les radiations électromagnétiques. Dans notre application, le bus CAN est un bus dédié qui assure une grande sécurité, seuls les capteurs étant connectés sur le bus. Le debit de transmission maximal est de 1Mo/s.

C.4.1 Données transmises par le bus CAN

Quand un événement sonore est détecté, une trame de données est envoyée sur le bus CAN par le PC d'analyse sonore.

La trame contient :

- le temps et la date de la détection (jour, mois, an, heure, minute, seconde et milliseconde)
- un drapeau indiquant le type d'événement sonore (parole ou son)
- un champ de caractères indiquant soit, les trois classes d'appartenance du son les plus probables, soit, l'expression de détresse identifiée, leurs vraisemblances et la localisation de l'événement sonore (la pièce).

Une trame sur le bus CAN contient 8 octets de données utiles et un identificateur qui représente le niveau de priorité. Les données du *capteur sonore intelligent* sont repartis en 41 trames et envoyés à travers du bus CAN. La séquence de transmission des données est représentée dans la figure C.4.

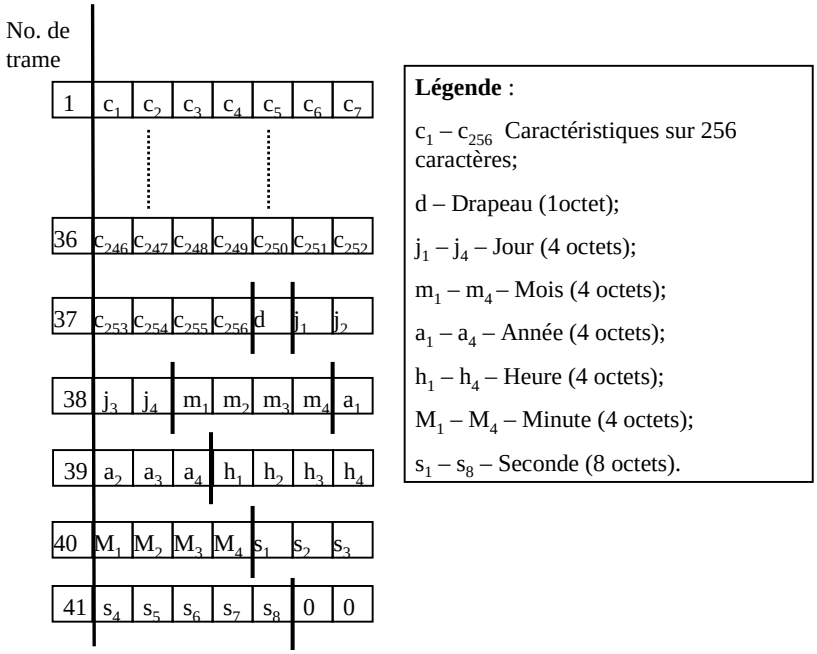


Fig. C.4: La structure des trames CAN pour la transmission des données du capteur sonore intelligent

Chapitre D

Format particuliers de fichiers

D.1 Fichiers d'étiquetage SAM

Les fichiers de type «SAM» sont utilisés dans les corpus de parole, [Standard SAM, 1992]. Ce sont des fichiers de type texte, [Wells et al., 1992]. Dans un corpus de parole plusieurs types de fichiers «SAM» sont nécessaires pour indiquer la composition du corpus, les locuteurs, etc.

Nous utilisons seulement le fichier «SAM» d'étiquetage d'un fichier sonore. Celui-là a un en-tête dans laquelle sont indiqués les informations suivantes :

TYP le type de fichier «SAM»

DIR le chemin du fichier SAM

SRC le chemin complet du fichier sonore correspondant

END le nombre d'échantillons du fichier sonore

EXP le nom de l'expert qui a crée le fichier

SYS le nom du logiciel utilisé pour créer le fichier

DAT la date de création

LBD le début de la zone des étiquettes

LBB l'étiquette qui est suivie de l'instant d'apparition exprimé en nombre d'échantillons depuis l'origine du fichier son, et, du nom de l'étiquette

ELF la fin du fichier

Pour la détection les étiquettes possibles sont :

Start Le début du signal

Stop La fin du signal

Détection L'instant de détection automatique

Ci-dessous, un exemple de fichier SAM est présenté :

```
LHD: V1.0
FIL: label
TYP: phonemic
VOL:
```

```
DIR: D:\fichiers_tests_detection\CD1\
SRC: D:\fichiers_tests_detection\CD1\0stouxx1blanc+20.wav
BEG: 0
END: 400000
EXP: generation automatique par logiciel
SYS: detection
DAT: 20.11.2001
SPA:
LBD:
LBB:    162080, ,   ,Start
LBB:    173040, ,   ,Stop
ELF:
```

D.2 Fichiers audio : WAV et SPH

Les fichiers sonores classiquement utilisés sous Windows sont de type «wav». Ce type de fichier est composé d'un en-tête et de la succession des échantillons du signal codés en binaire. L'en-tête contient, dans l'ordre, les informations suivantes codés en binaire :

- le type du fichier, chaîne «RIFF» codée sur 4 octets ASCII (identifie le standard multimedia Microsoft)
- la longueur en octets du fichier, codée sur 4 octets
- le mot «WAVE» codé sur 4 octets ASCII (identifie le type de fichier parmi ceux qui font partie du standard RIFF)
- le mot «fmt», codé sur 4 octets ASCII (précise le début de l'information de formattage)
- la longueur exprimée en octets de l'en-tête qui est fixé à 16, codée sur 4 octets
- le type de la quantification (linéaire ou logarithmique), habituellement PCM qui a le code 1, codé sur 2 octets
- le nombre des canaux sonores (mono=1, stereo=2), codé sur 2 octets
- la fréquence d'échantillonnage, codée sur 4 octets
- le nombre d'octets par seconde du son, codé sur 4 octets
- le nombre d'octets à envoyer simultanément à la carte de son, codé sur 2 octets
- le nombre de bits d'un échantillons (habituellement 16), codé sur 2 octets
- le mot «data» qui marque le début de la partie contenant les échantillons, codé sur 4 octets ASCII
- la longueur exprimée en octets de la partie des données, codée sur 4 octets

La plateforme ELISA utilise des fichiers de type SPHERE (.sph). La seule différence entre les formats Wave et SPHERE se situe au niveau de l'en-tête du fichier. L'en-tête d'un fichier SPHERE est en mode texte et elle contient une succession de champs codés en ASCII. Pour plus d'informations, voir référence [Format SPHERE, 1996]. Un exemple d'en-tête d'un fichier SPHERE est présenté ci-dessous :

```
NIST_1A
   1024
channel_count -i 1
sample_count -i 15866
sample_rate -i 16000
sample_n_bytes -i 2
sample_coding -s3 pcm
end_head
```

La première ligne est l'identification du système qui a créé le fichier et la deuxième la dimension de l'en-tête exprimée en nombre d'octets. La signification des champs est la suivante :

channel_count le nombre des canaux du fichier (mono = 1 / stéréo =2)

sample_count le nombre d'échantillons que compte le fichier

sample_rate la fréquence d'échantillonnage

sample_n_bytes le nombre d'octets d'un échantillon

sample_coding le type de codage, classiquement PCM (qui signifie que les échantillons du signal sont sans compression)

Le reste de l'en-tête est complété avec des caractères «\n» (0x0D) pour respecter la taille de l'en-tête figée à 1024 octets. Les échantillons du signal suivent l'en-tête et sont en format binaire.

Chapitre E

Logiciel de détection des événements sonores «HIS Detect»

L'implémentation en temps réel dans l'appartement d'étude de l'algorithme de détection d'événements sonores est réalisé avec un logiciel écrit sous LabWindows/CVI [N.I., 1999d]. LabWindows/CVI est un atelier de développement logiciel ANSI C, doté de fonctions graphiques, de fonctions du traitement du signal, de fonctions pour la commande des cartes d'acquisitions, etc [N.I., 1999e].

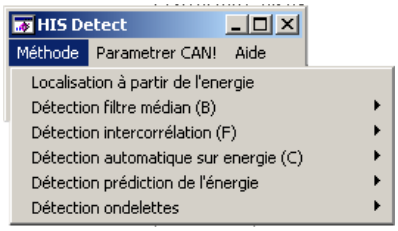


Fig. E.1: Logiciel de détection en temps réel sur N canaux (Menu Méthode)

Nous avons implémenté dans «HIS Detect», les meilleurs algorithmes de détection en temps réel sur les 5 canaux simultanément [N.I., 1999g]. Le menu «Méthode» (voir figure E.1) permet l'accès aux fonctions suivants :

Localisation à partir de l'énergie Permet la localisation du son par comparaison des énergies sur chacun des canaux

Détection filtre médian Réalise la détection d'événements sonores en temps réel sur 5 canaux avec l'algorithme utilisant un filtre médian conditionné.

Détection intercorrélation Réalise la détection d'événements sonores en temps réel indépendamment sur chaque des 5 canaux avec l'algorithme utilisant la fonction d'intercorrélation de deux fenêtres consécutives.

Détection prédiction de l'énergie Réalise la détection d'événements sonores en temps réel indépendamment sur chaque des 5 canaux avec l'algorithme fondé sur la prédiction de l'énergie par fonctions SPLINE.

Détection ondelettes Réalise la détection d'événements sonores en temps réel indépendamment sur chaque des 5 canaux avec l'algorithme utilisant la transformée en ondelettes.

L'instant de détection est envoyé au travers du bus CAN vers le PC «maître» quel que soit l'algorithme utilisé.

Le démarrage de la détection avec un des algorithmes proposés affiche l'écran présenté dans la figure E.2 [N.I., 1999b]. Nous affichons le signal du canal de la dernière détection, la position (la pièce) de la dernière détection sur un plan de l'appartement et une liste de toutes les détections [N.I., 1999c]. La liste des détections est enregistré dans un fichiers et les mêmes informations sont envoyés au PC Maître pour la fusion de données.

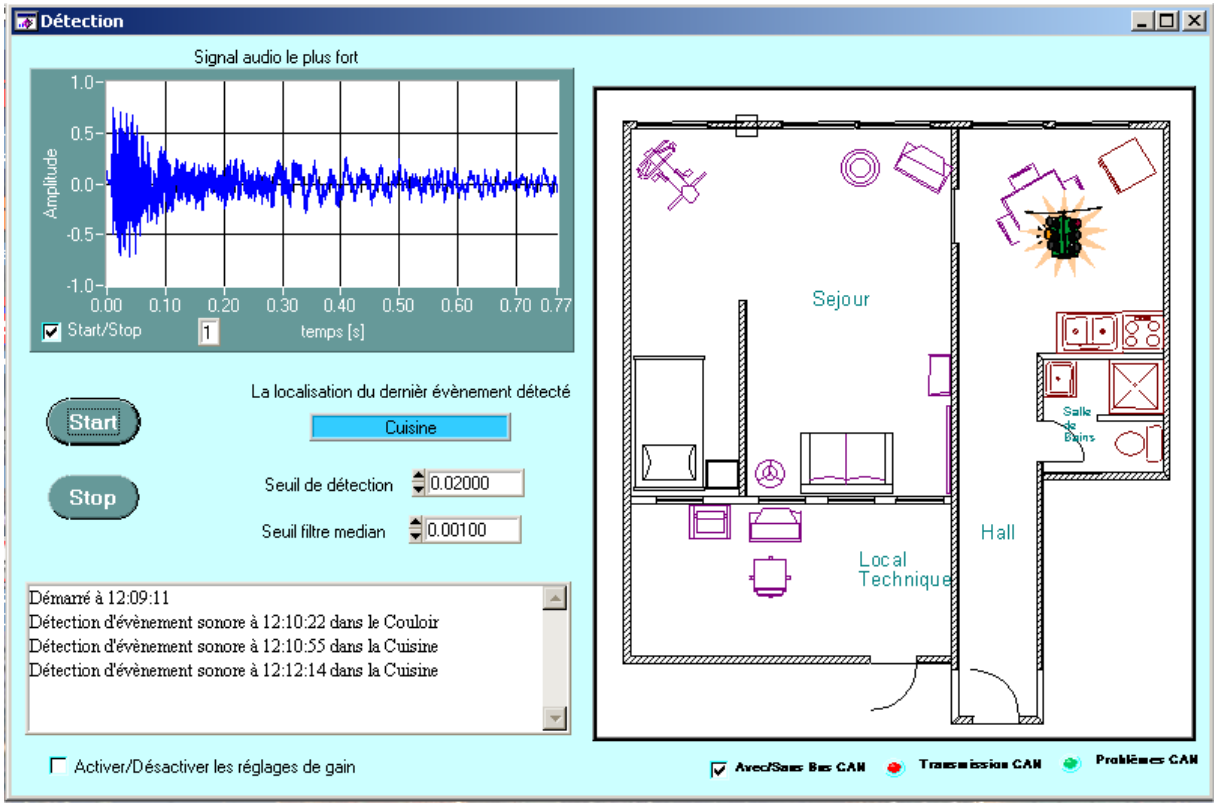


Fig. E.2: Logiciel de détection en temps réel sur N canaux (Affichage de la détection en temps réel)

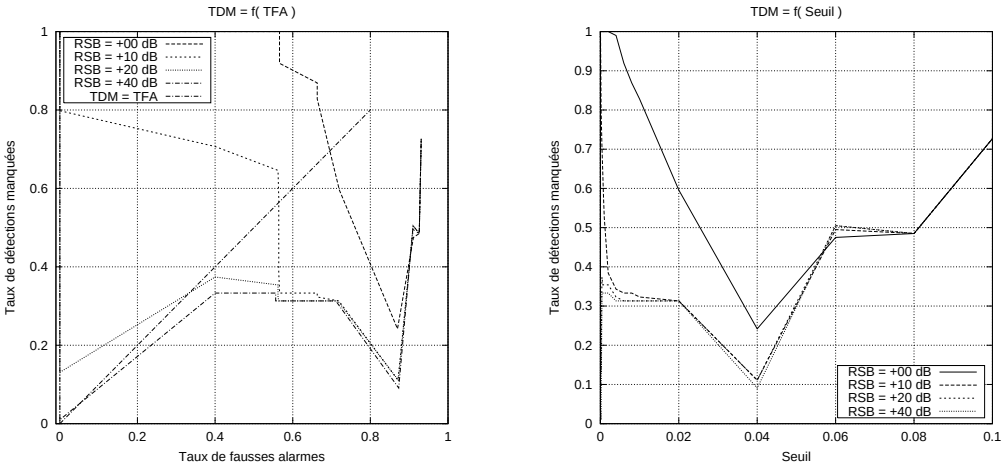
Chapitre F

Courbes ROC des algorithmes de détection

F.1 Les trois algorithmes issus de l'état de l'art

La courbe ROC[1] est la représentation du taux de détections manquées (TDM - voir définition (3.13) de la page 70) en fonction du taux de fausses alarmes (TFA - voir définition (3.14) de la page 70). Les courbes ROC pour les trois algorithmes de détection issus de l'état de l'art sont présentées dans les figures F.1, F.3 et F.4.

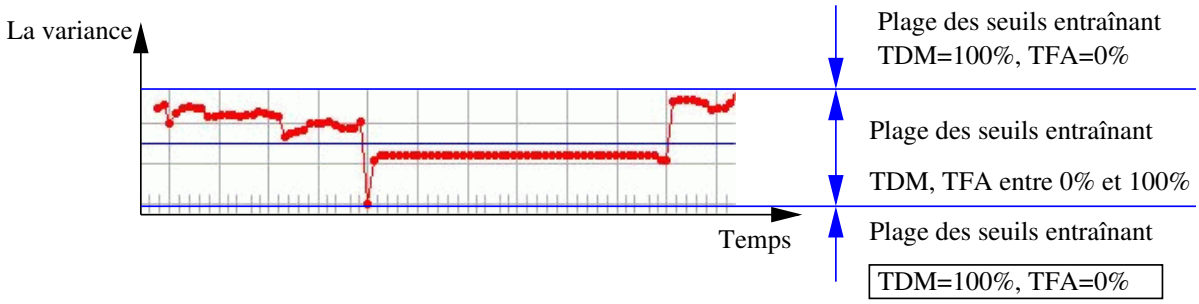
Algorithme fondé sur la variance : Sur sa courbe ROC (figure F.1(a)) nous observons qu'une valeur minimale du taux de détections manquées est obtenue pour un taux de fausses alarmes de 0.88%. Les grandes valeurs du taux de détection manquées de part et d'autre de ce point s'expliquent par le type particulier de seuillage utilisé : on considère qu'il y a détection si la variance descend en dessous du seuil .



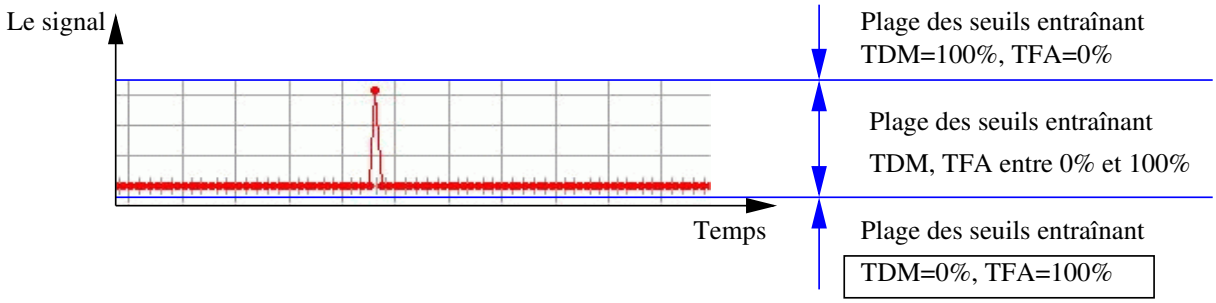
(a) Courbe ROC (b) TDM=f(seuil)

Fig. F.1: Courbe ROC et la courbe de TDM en fonction du seuil pour l'algorithme fondé sur la variance dans le cas du bruit HIS

[1] Receiver Operating Characteristics



(a) Détection si la variance descend en dessous du seuil



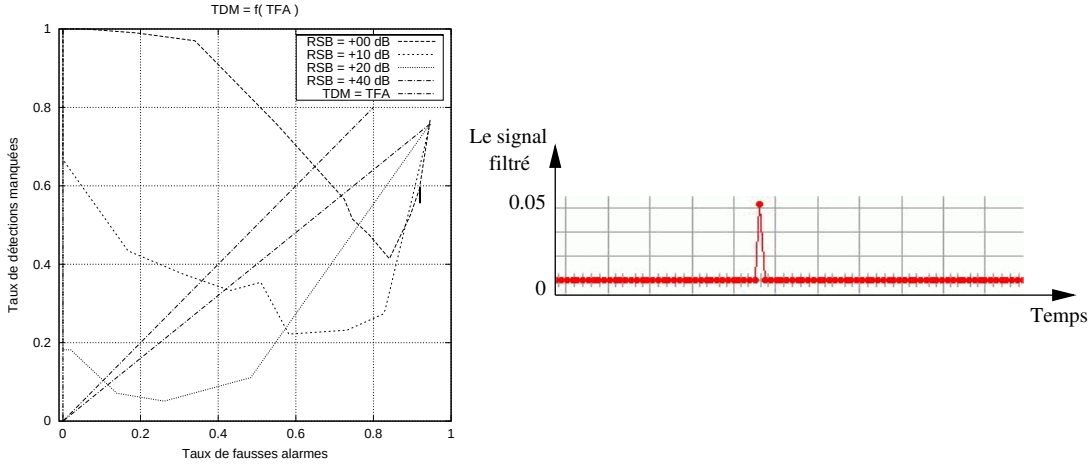
(b) Détection si la variance dépasse le seuil

Fig. F.2: Illustration de la variation des performances en fonction du seuil à partir de la représentation des évolutions de la variance au cours du temps

Dans le cas d'une détection classique par passage au dessus d'un seuil illustré à la figure F.2(b), lorsque ce seuil est proche de 0, aucune détection n'est manquée, TDM=0% ;lorsque le seuil augmente à partir d'un certain niveau il va y avoir de plus en plus de détections manquées. Le TDM va croître régulièrement de 0% à 100%.

Dans le cas illustré à la figure F.2(a) d'une détection par passage en dessous d'un seuil, rencontré dans la méthode de la variance, les résultats sont très différents. Lorsque le seuil est proche de 0, il y a 100% de détections manquées car la variance n'atteint jamais la valeur 0. Lorsque le seuil augmente, le nombre de détections s'accroît jusqu'à atteindre un palier à partir duquel il y aura diminution du nombre de détections, donc augmentation rapide du TDM. La variation du TDM en fonction du seuil prendra l'allure d'une courbe en «U» (voir figure F.1(b)).

Algorithme fondé sur le filtre médian : Le taux de fausses alarmes pour une valeur du RSB de 40 dB (voir figure F.3) varie brusquement de 0% à 100% parce que dans ce cas le signal filtré prend seulement deux valeurs (0 ou une valeur maximale).



(a) $TDM = f(TFA)$ (b) Illustration du signal filtré pour un RSB de 40 dB

Fig. F.3: Courbe ROC de l'algorithme fondé sur le filtre médian dans le cas du bruit HIS et illustration du signal filtré pour un RSB de 40 dB

Algorithme avec seuil adaptatif : Lorsque le TFA augmente et se rapproche de 100% le TDM augmente fortement à cause de l'inhibition de 2.5 secondes suivant toute détection(figure F.4), ce qui masque la détection d'un vrai événement lorsqu'il est trop proche.

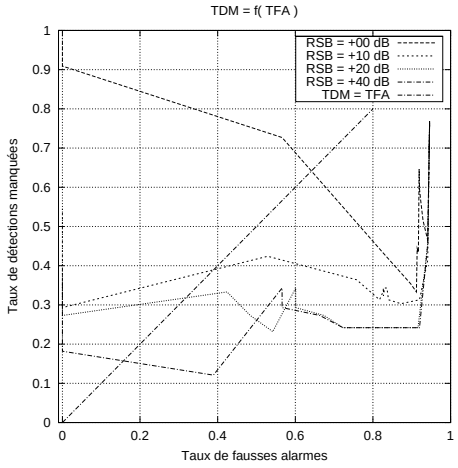


Fig. F.4: Courbe ROC de l'algorithme avec seuil adaptatif dans le cas du bruit HIS

F.2 Algorithme fondée sur la transformée en ondelettes

La courbe ROC de l'algorithme de détection fondée sur la transformée en ondelettes est visualisée sur la figure F.5. Les trois courbes ROC pour le RSB égal à 10 dB, 20 dB et respectivement 40 dB sont exactement superposés avec les axes du graphique (pour un taux de détection manquées égal à 0%, le taux de fausses alarmes est égal à 100% et inversement).

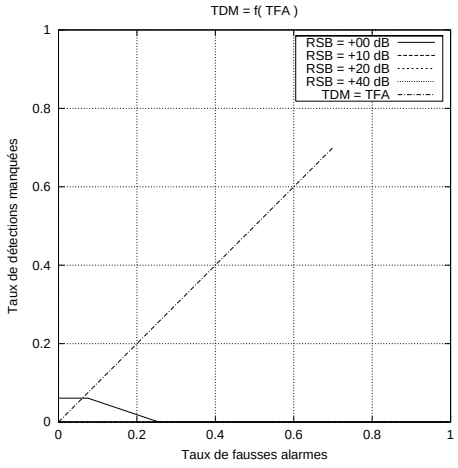


Fig. F.5: Courbe ROC de l'algorithme fondé sur la décomposition en ondelettes pour le bruit HIS

Chapitre G

Résultats détaillés de la classification sur la base de couplage

Le tableau G.1 présente la matrice de confusion de la classification pour la base de test de couplage avec un RSB variant entre 10 et 20 dB. Le nombre total de détections manquées est de 48 et celui des fausses alarmes est 205. Le TDM de la classification est proche de 3% et le TFA de la classification est de 11%.

	Classe reconnue / Classe en entrée	$\overline{A}$				A		
		C1	C3	C4	C9	C2	C5	C7
$\overline{A}$	Claquements de porte (C1)	**399**	**0**	**70**	**0**	**51**	**0**	**3**
	Sonneries de téléphone (C3)	**69**	**331**	**0**	**0**	**0**	**20**	**96**
	Sons de pas (C4)	**1**	**0**	**12**	**0**	**0**	**0**	**0**
	Serrures de porte (C9)	**20**	**0**	**1**	**144**	**35**	**0**	**0**
A	Bris de verre (C2)	**1**	**0**	**0**	**0**	**87**	**0**	**0**
	Cris (C5)	**3**	**2**	**0**	**0**	**0**	**63**	**5**
	Vaisselle (C7)	**36**	**0**	**6**	**0**	**15**	**0**	**106**

Nombre Total de sons=1614, DM_{Total} = 48, FA_{Total} = 205

$TDM_{Classification}$ = 2.9%

$TFA_{Classification}$ = 11.2%

Tab. G.1: Matrice de confusion en nombre de fichiers pour la classification des sons de la base de couplage (RSB compris entre 10 et 20 dB)

Le tableau G.2 présente la matrice de confusion de la classification pour la base de test de couplage avec un RSB qui varie entre 0 et 40 dB. Dans ce cas le nombre de détections manquées est de 46 et le nombre de fausses alarmes est de 188.

Donc dans les deux cas que nous venons d'étudier, correspondant à des plages du rapport signal sur bruit de 10 à 20 dB et de 0 à 40 dB, les résultats sont très voisins aussi bien du point de vue du TDM que du TFA.

	Classe reconnue / Classe en entrée	$\overline{A}$				A		
		C1	C3	C4	C9	C2	C5	C7
$\overline{A}$	Claquements de porte (C1)	**393**	**0**	**73**	**0**	**55**	**0**	**2**
	Sonneries de téléphone (C3)	**54**	**371**	**0**	**0**	**0**	**12**	**80**
	Sons de pas (C4)	**3**	**0**	**10**	**0**	**0**	**0**	**0**
	Serrures de porte (C9)	**10**	**0**	**4**	**147**	**38**	**0**	**1**
A	Bris de verre (C2)	**0**	**0**	**0**	**1**	**87**	**0**	**0**
	Cris (C5)	**2**	**2**	**1**	**0**	**0**	**64**	**4**
	Vaisselle (C7)	**34**	**1**	**5**	**0**	**17**	**2**	**104**

Nombre Total de sons=1546, $DM_{Total} = 46$, $FA_{Total} = 188$

$TDM_{Classification} = 2.9\%$

$TFA_{Classification} = 10.8\%$

Tab. G.2: Matrice de confusion en nombre de fichiers pour la classification des sons de la base de couplage (RSB compris entre 0 et 40 dB)

Chapitre H

Liste des publications de l'auteur

Le travail décrit dans ce mémoire a été présenté dans les articles ci-dessous :

Conférences Internationales avec comité de lecture et actes

- E.Castelli, D.Istrate, V.Rialle, N.Noury, «Information extraction from speech in stress situation. Application to the Medical Supervision in a Smart House», *Conférence ORAGE (ORAlité et GEstualité)* , Aix-en-Provence, France, 18-22 juin 2001, pp. 362-371

- E.Castelli and D.Istrate, «Multichannel Audio Acquisition for Medical Supervision in an Intelligent Habitat», *European Conference on Circuits Theory and Devices (ECCTD'01)*, Espoo/Helsinki, Finlande, 28-31 Août 2001, , pp. II-1 II-4

- D.Istrate and E.Castelli, «Everyday Life Sounds and Speech Analysis for a Medical Telemonitoring System», *EUROSPEECH Conference*, Aalborg, Danemark, 3-7 septembre 2001,pp. E15 2417-2420

- D.Istrate and E.Castelli, «Multichannel Sound Acquisition with Stress Situations Determination for Medical Supervision in a Smart House», *Text Speech and Dialogue Conference*, Zelezna Ruda, République Tchèque, 10-13 septembre 2001, pp. 266-272

- M.Vacher, D.Istrate, L.Besacier, J.F.Serignat and E.Castelli, «Smart Audio Sensor for Telemedicine», *Smart Objects Conferences sOc'2003*, Grenoble, France, 15-17 Mai 2003, pp. 222-225

- E.Castelli, M.Vacher, D.Istrate, L.Besacier and J.F.Serignat, «Habitat Telemonitoring System Based on the Sound Surveillance», *International Conference on Information Communication Technologies in Health*, Samos, Greece, 13-15 Juillet 2003, pp. 141-146

- M.Vacher, D.Istrate, L.Besacier, J.F.Serignat and E.Castelli, «Life Sounds Extraction and Classification in Noisy Environment», *International Association of Science and TEchnology for Development Conference on Signal Image Processing*, Honolulu,Hawaii, 13-15

Août 2003

- D. Istrate, G. Virone, M. Vacher, E. Castelli, J. F. Serignat, «Communication Between A Multichannel Audio Acquisition And An Information System In A Health Smart Home For Data Fusion», *International Association of Science and TEchnology for Development Conference on Internet and Multimedia Systems and Networks*, Honolulu,Hawaii, 13-15 Août 2003

- G. Virone, D. Istrate, M. Vacher, J. F. Serignat, N. Noury et J. Demongeot, «First Steps in Data Fusion between a Multichannel Audio Acquisition and an Information System for Home Healthcare», *IEEE Engineering In Medicine And Biology Society Conference*, Cancun, Mexique, 13-15 Septembre 2003, pp. 1364-1367

- M. Vacher, D. Istrate, L. Besacier, J.F.Serignat et E. Castelli, «Sound Detection and Classification for Medical Telesurvey», *IASTED International Conference on Biomedical Engineering*, Innsbruck, Autriche

Forum étudiant ICASSP 2001

- D.Istrate and Q.C.Nguyen, «Room Echo Cancellation for Speech Recognition», *ICASSP¨2001 (International Conference on Acoustics, Speech and Signal Processing)* - Student Forum, Salt Lake City, Utah, USA, 6-11 Mai 2001

- Q.C.Nguyen, D.Istrate and J.Barton, «Blind Source Separation», *ICASSP 2001 (International Conference on Acoustics, Speech and Signal Processing)* - Student Forum, Salt Lake City, Utah, USA, 6-11 Mai 2001

Rapports techniques internes

- D. Istrate and M.Vacher, «Détection de signaux sonores noyés dans le bruit», février 2003, CLIPS-IMAG, Grenoble, France

- D. Istrate and M.Vacher, «Reconnaissance des classes de sons», juin 2003, CLIPS-IMAG, Grenoble, France

- D. Istrate and M.Vacher, «Application des ondelettes à la détection de signaux sonores noyés dans le bruit», juillet 2003, CLIPS-IMAG, Grenoble, France

Chapitre I

Abréviations

CMS	**C**epstral **M**ean **S**ubtraction
DMA	**D**irect **M**emory **A**ccess
DTW	**D**irect **T**ime **W**arping
DWT	**D**irect **W**avelet **T**ransform
DWTC	**D**irect **W**avelet **T**ransform based **C**oefficients
EM	**E**xpectation **M**aximization
FFT	**F**ast **F**ourier **T**ransform
GAET	**G**eometrically **A**daptative **E**nergy **T**hreshold
GMM	**G**aussien **M**ixture **M**odel
HIS	**H**abitat **I**ntelligent **S**anté
HMM	**H**idden **M**arkov **M**odel
LDA	**L**inear **D**iscrimination **A**nalysis
LFCC	**L**inear-**F**requencies **C**epstral **C**oefficients
LPC	**L**inear **P**rediction **C**oefficients
LPCC	**L**inear **P**rediction **C**epstral **C**oefficients
LSB	**L**ast **S**ignifiant **B**it
LSPE	**L**east-**S**quare **P**eriodicity **E**stimator
MFCC	**M**el-**F**requencies **C**epstral **C**oefficients
PCA	**P**rincipal **C**omponent **A**nalysis
PCM	**P**ulse **C**oded **M**odulation
ROC	**R**eceiver **O**perating **C**haracteristics
RSB	**R**apport **S**ignal sur **B**ruit
SONAR	**SO**und **NA**vigation and **R**anging
TDM	**T**aux de **D**étections **M**anquées
TEE	**T**aux d'**E**gale **E**rreur
TFA	**T**aux de **F**ausses **A**larmes

TFD	**T**ransformée de **F**ourier **D**iscrète
TFR	**T**ransformée de **F**ourier **R**apide
VAD	**V**oice **A**ctivity **D**etection
VQ	**V**ector **Q**uantization
ZCR	**Z**ero **C**rossing **R**ate

Bibliographie

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automation Control*, 19 :716–723.

[Antoniadis, 2003] Antoniadis, A. (2003). Compression et débruitage avec les ondelettes. Technical report, LMC - IMAG.

[Bellanger, 2002] Bellanger, M. (2002). *Traitement Numérique du Signal. Théorie et Pratique*. ISBN 2-10-006311-1. Dunod, Paris, 7ème edition.

[Boite et al., 2000] Boite, R., Bourlard, H., Dutoit, T., Hang, J., and Leich, H. (2000). *Traitement de la parole*. ISBN 2-88074-388-5. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Boor, 1978] Boor, C. D. (1978). *A Practical Guides to SPLINES*. Springer-Verlag, New York.

[Bouvet, 1992] Bouvet, M. (1992). *Traitement des signaux pour les systèmes sonar*. ISBN 2-225-82615-3. Masson, Paris.

[Cabell and Fuller, 1989] Cabell, R. H. and Fuller, C. R. (1989). A smart pattern recognition system for the automatic identification of aerospace acoustic sources. *AIAA 12th Aeroacoustics Conference*, pages 321–325, San Antonio, USA.

[CALLIOPE, 1989] CALLIOPE (1989). *La parole et son traitement automatique*. ISBN 2-225-81516-X. Masson, Paris.

[CAN Bus site, 2003] CAN Bus site (2003). CAN Bus. http ://www.can.bosch.com.

[Cappé, 2001] Cappé, O. (2001). Modèles de mélange et modèles de markov cachés pour le traitement automatique de la parole. *[http ://tsi.enst.fr/ cappe/h2m/index.html]*, (ENST/Paris) :1–9.

[Castelli et al., 2002] Castelli, E., Serignat, J., and Rialle, V. (2002). Rapport final du projet RESIDE-HIS (Reconnaissance de situations de détresse en Habitat Intelligent Santé). Technical report, CLIPS et TIMC.

[Chagnolleau et al., 2001] Chagnolleau, I. M., Gravier, G., and Blouet, R. (2001). Overview of the ELISA consortium research activities. *2001 : a Speaker Odyssey*, (2) :67–72.

[Chang and Wu, 2000] Chang, D. C. and Wu, W. R. (2000). Feedback median filter for robust preprocessing of glint noise. *IEEE Transactions on Aerospace and Electronic Systems*, 22(6) :213–221.

[Chevrolet et al., 2002] Chevrolet, J. C., Denz, M., Merminod, B., Osswald, S., and Roulet, M. (2002). Télémédecine CH. Technical report, Académie Suisse des Sciences Médicales.

[Chiu et al., 2001] Chiu, P., Boreczky, J., Girgensohn, A., and Kimber, D. (2001). Liteminutes : An Internet-Based system for multimedia meeting minutes. *World Wide Web Conference*, pages 140–149, Hong-Kong.

[Colin, 2002] Colin, J. M. (2002). *Le Radar. Théorie et pratique*. ISBN 2-7298-1176-1. Ellipses,Paris.

[Couvreur, 1997] Couvreur, C. (1997). *Environmental Sound Recognition : A statistical approach*. PhD thesis, Faculté Polytechnique de Mons, Belgique.

[Cowling and Sitte, 2002] Cowling, M. and Sitte, R. (2002). Analysis of non speech recognition techniques for use in a non-speech sound recognition system. *6th International Symposium on Digital Signal Processing for Communication Systems - Sydney-Manly, Australie*, page 4pages.

[Cutler, 2003] Cutler, R. (2003). The distributed meetings system. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 756–759, Hong-Kong.

[Davis and Mermelstein, 1980] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Audio, Speech and Signal Processing*, 28(4) :357–366.

[Dubois, 2001] Dubois, E. (2001). *Chirurgie augmentée, un cas de réalité augmentée ; Conception et réalisation centrées sur l'utilisateur*. PhD thesis, Université Joseph Fourier, Grenoble.

[Dufaux, 2001] Dufaux, A. (2001). *Detection and recognition of Impulsive Sounds Signals*. PhD thesis, Faculté des sciences de l'Université de Neuchâtel, Suisse.

[Dufaux et al., 2000] Dufaux, A., Besacier, L., Ansorge, M., and Pellandini, F. (2000). Automatic sound detection and recognition for noisy environment. *European Signal Processing Conference (EUSIPCO), Tampere, Finlande*, pages 1033–1036.

[Duvaut, 1991] Duvaut, P. (1991). *Traitement du signal*. ISBN 2-86601-422-7. Hermès, Paris, 2ème edition.

[El-Maleh et al., 2000] El-Maleh, K., Klein, M., Petrucci, G., and Kubal, P. (2000). Speech/music discrimination for multimedia applications. *International Conference on Acoustics, Sound and Signal Processing (ICASSP), Istanbul, République Turque*.

[Fieschi, 2000] Fieschi, M. (2000). Les médecins aussi tissent leur toile. *Supplément La Recherche*, pages 15–19.

[Format SPHERE, 1996] Format SPHERE (1996). Format SPHERE. ftp ://jaguar.ncsl.nist.gov/pub/sphere_2.6a.tar.Z.

[G. Tzanetakis and Cook, 2001] G. Tzanetakis, G. E. and Cook, P. (2001). Audio analysis using the discrete wavelet transform. *WSES Int. Conf. Acoustics and Music : Theory and Applications (AMTA 2001)*, pages 225–229, Skiathos, Greece.

[Gargour and Samir, 2001] Gargour, C. S. and Samir, C. (2001). *Traitement numérique des signaux*. ISBN 2-921145-24-3. Ecole de technologie supérieur, Canada.

[Gauvain et al., 1999] Gauvain, J., Lamel, L., and Adda, G. (1999). Audio partitionning and transcription for broadcast data indexation. *First European Workshop on Content-Based Multimedia Indexing CBMI'99*, pages 67–73, Toulouse, France.

[Gökhun and Özer, 2000] Gökhun, S. and Özer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 8(4) :478–482.

[Goldhor, 1993] Goldhor, R. S. (1993). Recognition of environmental sounds. *International Conference on Audio, Speech and Signal Processing*, volume 1, pages 149–152, Minneapolis, USA.

[Goldstein, 1976] Goldstein, U. (1976). Speaker-identifying features based on formant tracks. *Journal of the Acoustical Society of America*, 59(1) :176–182.

[H. Hacihabiboglu and Nishan, 2002] H. Hacihabiboglu, F. H. and Nishan, C. (2002). Musical instrument recognition with wavelet envelopes. *EAA Convention, Proc. Forum Acusticum Sevilla*, pages 356–360, Sevilla, Spain.

[Häcker et al., 2002] Häcker, J., Engelhardt, F., and Frey, D. D. (2002). Robust manufacturing inspection and classification with machine vision. *International Journal of Production Research*, 40(6) :1319–1334.

[Irwin, 1980] Irwin, M. J. (1980). Periodicity estimation in the presence of noise. *Acoustics Conference '80*, pages 243–247, Windemere, U.K.

[JAIN et al., 2000] JAIN, A. K., DIUN, R. P. W., and MOA, J. (2000). Statistical pattern recognition : A review. *IEEE Trans. PAMI*, 22(1) :4–37.

[Jolliffe, 1986] Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

[Jouvet, 1988] Jouvet, D. (1988). *Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris.

[Junqua et al., 1991] Junqua, J. C., Reaves, B., and Mak, B. (1991). A study of endpoint detection algorithms in adverse conditions : Incidence on a DTW and HMM recognize. *Eurospeech '91*, pages 1371–1374, Genova, Italie.

[Kil and Shin, 1996] Kil, D. and Shin, F. (1996). *Pattern Recognition and Prediction with Applications to Signal Characterization*. AIP Press Woodbury, New York.

[Kintzig et al., 2002] Kintzig, C., Poulain, G., Privat, G., and Favennec, P. N. (2002). *Objets communicants*. ISBN 2-7462-0475-4. LAVOISIER Hermès Science Publications, Paris.

[Kirsteins et al., 1997] Kirsteins, I. P., Mehta, S. K., and Fay, J. (1997). Power-law processors for detecting unkown signals in colored noise. *International Conference on Acoustics, Sounds and Signal Processing*, page 4pages, Munich, Allemagne.

[Kornblum et al., 2001] Kornblum, C., Sibony, O., Rol, A. L., Strauss, A., Champetier, D., Lavictoire, M., and Berry, M. (2001). Télémédecine & industrialisation. Technical report, Ministère de l'Emploi et de la Solidarité.

[Kunt et al., 1991] Kunt, M., Bellanger, M., et al. (1991). *Techniques modernes de traitement numérique des signaux*, volume 1-3 of *ISBN 2-88074-207-2*. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Kunt et al., 2000] Kunt, M., Coray, G., Granlund, G., Haton, J., Ingold, R., and Kocher, M. (2000). *Reconnaissance des formes et analyse de scènes*, volume 1-3 of *ISBN 2-88074-384-2*. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Lacoume, 1997] Lacoume, J. L. (1997). *Statistiques d'ordre supérieur pour le traitement du signal*. ISBN 2-225-83118-1. Masson, Paris.

[Lau et al., 2002] Lau, C., Churchill, R. S., Kim, J., et al. (2002). Asynchronous web-based patient-centered home telemedicine system. *IEEE Transactions on Biomedical Engineering*, 49(12) :1452–1459.

[Lee et al., 1994] Lee, W., Kim, Y., Gove, R. J., and Read, C. J. (1994). MediaStation 5000 : Integrating video and audio. *IEEE Multimedia*, 1(2) :50–61.

[Linde et al., 1980] Linde, Y., Buzo, A., and Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, COM-28 :84–95.

[LITMED Project, 2002] LITMED Project (2002). Litmed II Project. http ://www.litmed.net/.

[Mallat, 2000] Mallat, S. (2000). *Une exploration des signaux en ondelettes*. ISBN 2-7302-0733-3. Les Editions de l'Ecole Polytechnique, Paris.

[Mallat and Hwang, 1991] Mallat, S. and Hwang, W. L. (1991). Singularity detection and processing with wavelets. Technical report, Courant Institute of Mathematical Sciences, New York University.

[Max, 1989] Max, J. (1989). *Traitement du Signal et Applications aux Mesures Physiques*, volume 1-4 of *ISBN 2-225-80470-2*. Masson, Paris, 4ème edition.

[McCowan et al., 2003] McCowan, I., Bengio, S., Gatica-Perez, D., et al. (2003). Modeling human interaction in meetings. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 748–751, Hong-Kong.

[Meignier et al., 2000] Meignier, S., Bonastre, J., Fredouille, C., and Merlin, T. (2000). Evolutive HMM for multi-speaker tracking system. *International Conference on Audio, Speech and Signal Processing ICASSP '2000*, pages 543–547, Istanbul, République Turque.

[Morgan et al., 2003] Morgan, N., Baron, D., Bhagat, S., et al. (2003). Meetings about meetings : Research at ICSI on speech in multiparty conversations. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 740–743, Hong-Kong.

[Myers and Rabiner, 1981] Myers, C. S. and Rabiner, L. R. (1981). A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60(7) :1389–1409.

[Nemer et al., 2001] Nemer, E., Goubran, R., and Mahmoud, S. (2001). Robust voice activity detection using higher-order statistics in the lpc residual domain. *IEEE Transactions on Speech and Audio Processing*, 9(3) :217–231.

[N.I., 1999a] N.I. (1999a). *6034E/6035E User Manual*. National Instruments Corporation.

[N.I., 1999b] N.I. (1999b). *LabWindows/CVI Programmer Reference Manual*. National Instruments Corporation.

[N.I., 1999c] N.I. (1999c). *LabWindows/CVI User Interface Reference Manual*. National Instruments Corporation.

[N.I., 1999d] N.I. (1999d). *LabWindows/CVI User Manual*. National Instruments Corporation.

[N.I., 1999e] N.I. (1999e). *Multithreading in CVI.* National Instruments Corporation.

[N.I., 1999f] N.I. (1999f). *PCI E Series User Manual*. National Instruments Corporation.

[N.I., 1999g] N.I. (1999g). *Standard Libraries Reference Manual*. National Instruments Corporation.

[Olsen et al., 2003] Olsen, B. I., Eggen, A. E., Bellika, J. G., et al. (2003). An electronic health record-based system for automatic monitoring and control of medication. *International Conference on Information Communication Technologies in Health, Samos, Greece*, pages 147–151.

[Ozer and Tanyer, 1998] Ozer, H. and Tanyer, S. G. (1998). A geometric algorithm for voice activity detection in nonstationary gaussian noise. *EUropean SIgnal Processing COnference '98*, pages 543–547, Rhodes, Grèce.

[Papadopoulos et al., 1992] Papadopoulos, G., Efstathiou, K., Li, Y., and Delis, A. (1992). Implementation of an intelligent instrument for passive recognition and two-dimensional location estimation of acoustic targets. *IEEE Transactions On Instrumentation and Measurement*, 41(6) :885–890.

[Paradie and Nawab, 1990] Paradie, M. J. and Nawab, S. H. (1990). The classification of ringing sounds. *International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 2435–2438, New Mexico, USA.

[Partnership, 2001] Partnership, R. W. C. (1998-2001). CD - Sound scene database in real acoustical environments. http ://tosa.mri.co.jp/sounddb/indexe.htm.

[Pfeiffer et al., 1996] Pfeiffer, S., Fischer, S., and Effelsberg, W. (1996). Automatic audio content analysis. *Reihe Informatik*, 8 :15–27.

[Press et al., 2002] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (2002). *Numerical Recipes in C ; The Art of scientific Computing ;The second Edition*. ISBN 0-521-43108-5. Cambridge University Press.

[Prieur, 2003] Prieur, C. (2003). L'été le plus meurtrier en france. *Le Monde, 9 Septembre*.

[Projet RESIDE-HIS, 2000] Projet RESIDE-HIS (2000). Projet RESIDE-HIS. http ://www-clips.imag.fr/tech-adm/perso/michel.vacher/HIS/SITE_WEB/index.htm.

[Projet TIISSAD, 2001] Projet TIISSAD (2001). TIISSAD project. http ://www.loria.fr-/projets/TIISSAD/.

[PROTEL, 1999] PROTEL (1999). *Protel 99 SE Handbook*.

[Équipe GEOD Dan Istrate, 2001] Équipe GEOD Dan Istrate, C.-I. (2001). Base de données. Sons de la vie courante.

[Rabiner and Juang, 1993] Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of speech recognition*. ISBN 0-13-015157-2. Prentice Hall PTR, New Jersey, USA.

[Rabiner and Sambru, 1975] Rabiner, L. R. and Sambru, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *Bell Systems Technical Journal*, 54(2) :297–315.

[Renals and Ellis, 2003] Renals, S. and Ellis, D. (2003). Audio information access from meeting rooms. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 744–747, Hong-Kong.

[Reynolds, 1994] Reynolds, D. (1994). Speaker identification and verification using gaussian mixture speaker models. *Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Suisse*, pages 27–30.

[Rialle et al., 1999] Rialle, V., Lauvernay, N., Franco, A., Piquard, J. F., and Couturier, P. (1999). A smart room for hospitalized elderly people : Essay of modelling and first steps of an experiment. *Technology and Health Care*, 7 :343–357.

[Roeder and Wasserman, 1997] Roeder, K. and Wasserman, L. (1997). Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92 :894–902.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464.

[Sciascia, 1992] Sciascia, S. (1992). CD - bruitages, vol.3.

[Scott et al., 1993] Scott, E. A., Fuller, C. R., O'Brien, W. F., and Cabell, R. H. (1993). Sparse distributed associative memory for the identification of aerospace acoustic sources. *AIAA Journal*, 31(9) :1583–1589.

[Seck, 2001] Seck, M. (2001). *Détection de ruptures et suivi de classe de sons pour l'indexation sonore*. PhD thesis, Université de Rennes I.

[Seck et al., 2001] Seck, M., Magrin-Chagnolleau, I., and Bimbot, F. (2001). Experiments on speech tracking in audio documents using gaussian mixture modeling. *International Conference on Acoustics, Sounds and Signal Processing*, Salt Lake City, Utah, USA.

[Sohn et al., 1999] Sohn, J., Kim, N., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1) :1–3.

[Standard SAM, 1992] Standard SAM (1992). http ://www.icp.grenet.fr/relator/standsam.html.

[Stanford et al., 2003] Stanford, V., Garofolo, J., Galibert, O., Michel, M., and Laprun, C. (2003). The NIST smart space and meeting room projects : Signals, acquisition, annotation and metrics. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 736–739, Hong-Kong.

[Sundaram and Chang, 2000] Sundaram, H. and Chang, S. F. (2000). Audio scene segmentation using multiple features, models and time scales. *ICASSP, Istanbul, République Turque*.

[Tanyer and Ozer, 2000] Tanyer, S. G. and Ozer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 8(4) :478–482.

[Tepedelenlioglu et al., 2001] Tepedelenlioglu, C., Sidiropoulos, N., and Giannakis, G. (2001). Median filtering for power estimation in mobile communication systems. *IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications, Taoyouan, Taiwan*, pages 229–231.

[Theodoridis and Koutroumbas, 1998] Theodoridis, S. and Koutroumbas, K. (1998). *Pattern Recognition*. Academic Press, San Diego.

[Truchetet, 1998] Truchetet, F. (1998). *Ondelettes pour le signal numérique*. ISBN 2-86601-672-6. Hermes.

[UIUC Project, 2002] UIUC Project (2002). UIUC Dysphagia project. http ://white.shs.uiuc.-edu :8080/.

[Valens, 1999] Valens, C. (1999). *A Really Friendly Guide to Wavelets*. http ://perso.wanadoo.fr/polyvalens/clemens/wavelets/wavelets.html.

[Vassiliadis, 2003] Vassiliadis, K. (2003). DROMEAS - a wearable platform for the monitoring of health condition and sport performance of athletes and the real-time prevention of sport injuries. *International Conference on Information Communication Technologies in Health, Samos, Greece*, pages 136–140.

[VBCH Project, 2002] VBCH Project (2002). Van Buren County Hospital project. http ://showcase.netins.net/web/forhealth/telemed.htm.

[Virone et al., 2002] Virone, G., Noury, N., and Demongeot, J. (2002). A system for automatic measurement of circadian activity in telemedicine. *IEEE Transactions on Biomedical Engineering*, 49(12) :1463–1469.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, IT-13 :260–269.

[Waibel et al., 2003] Waibel, A., Schultz, T., Bett, M., et al. (2003). Smart : The smart meeting room task at ISL. *International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 752–755, Hong-Kong.

[Wells et al., 1992] Wells, D., Barry, J., Grice, W., Fourcin, M., and Gibbon, A. (1992). SAM ESPRIT PROJECT 2589 - multilingual speech input/output assessment, methodology and standardisation. Final report. Technical Report SAM-UCL-G004, University CollegeLondon.

[Woodard, 1992] Woodard, J. P. (1992). Modeling and classification of natural sounds by product code hidden markov models. *IEEE Transactions on Signal Processing*, 40(7) :1833–1835.

[Wren et al., 1997] Wren, C. R., Sparacino, F., et al. (1997). Perceptive spaces for performance and entertainment : Untethered interaction using computer vision and audition. *Applied Artificial Intelligence (AAI) Journal*, pages 267–284.

[Yamada and Watanabe, 2001] Yamada, T. and Watanabe, N. (2001). Voice activity detection using non-speech models and HMM composition. *Workshop on Hands-free Speech Communication, Tokyo, Japan*, pages 323–327.

[Yaniv and Burshtein, 2003] Yaniv, R. and Burshtein, D. (2003). An enhanced dynamic time warping model for improved estimation of DTW parameters. *IEEE Transactions on Speech and Audio Processing*, 11(3) :216–228.

[Zhang et al., 2002] Zhang, J., Ward, W., and Pellom, B. (2002). Phone based activity detection using online bayesian adaptation with conjugate normal distributions. *International Conference on Acoustics, Sounds and Signal Processing*, pages 123–127, Orlando, Floride, USA.

Index

Résumé

Depuis quelques années se développe le concept général d'espace perceptif ou salle intelligente qui répond de diverses façons aux besoins, demandes, attentes des acteurs humains. Ce travail de thèse se situe à la frontière entre les espaces perceptifs et la télémédecine. Dans ce contexte, cette thèse analyse et propose des solutions aux problématiques spécifiques au traitement du son dans les espaces perceptifs plus particulièrement pour la télésurveillance médicale. Parmi ces problématiques la classification automatique de sons de la vie courante a été très peu explorée jusqu'à aujourd'hui. Dans ce travail, un système d'analyse sonore en deux étapes est proposé pour éviter d'analyser un flux audio continu. Le rôle de la détection des événements sonores est d'extraire du bruit environnemental les signaux à identifier. Les algorithmes issus de l'état de l'art se montrant insuffisamment efficaces dans nos conditions, de nouveaux algorithmes mieux adaptés aux signaux impulsionnels, comme ceux utilisant la transformée en ondelettes sont proposés. Pour la classification des sons proprement dite, l'utilisation de techniques issues de la reconnaissance automatique de la parole est d'abord envisagée. Ces techniques sont ensuite enrichies par l'ajout de paramètres acoustiques mieux adaptés, parmi lesquels ceux issus de la transformée en ondelettes et de la détection de signaux musicaux. Les performances de la classification sont aussi évaluées dans le bruit et une solution de pré-traitement est présentée. Les problématiques liées au couplage entre la détection et la classification, ainsi que le problème de l'évaluation d'un tel système sont aussi abordées dans ce travail. Une implémentation en temps réel des algorithmes proposés a été réalisée pour l'application de télésurveillance médicale et est en cours de validation dans l'appartement test disponible pour le projet.

Mots-clés : détection, classification, GMM, ondelettes, paramètres acoustiques, télémédecine, espaces perceptifs, traitement du signal.

Abstract

From few years, the general concept of perceptive spaces or smart rooms that answers in different way to the human actors needs, demands or expectations is in a continuous developing. The work presented in this thesis is set on the border of the perceptive spaces and telemedecine. Thus, this thesis analyzes and proposes solutions to sound processing problems for the perceptives spaces, generally and for medical telemonitoring, particulary. From all the problems linked to perceptive spaces, the automatic classification of everyday life sounds has not been explored too much until now. In order to avoid the classification of a continuous audio flow, a two steps sound analysis system is proposed in this work. The role of the sound event detection, the first step of the proposed system, is to extract the signal to be identified from the environmental noise. The state of the art algorithms are not efficient enough in our work conditions and thus, new algorithms better adapted to impulsive signals like those using the wavelet transform are proposed. Concerning the sound classification itself, the second step of the proposed system, a first approach was the use of the automatic speech recognition techniques. These techniques are improved by adding better adapted acoustical parameters among which those determined with wavelet transform and those used to detect the musical signals. The performances of the classification method are determined in a noisy environment and a preprocessing solution is presented too. The problems concerning the coupling of detection and classification steps, as well as the system evaluation are presented. A real-time implementation of the proposed algorithms was realized for a medical telemonitoring application which is in a validation process in our test apartment.

Key words : detection, classification, GMM, wavelets, acoustical parameters, telemedicine, perceptive spaces, signal processing.