Caractérisation et reconnaissance de sons d'eau pour le suivi des activités de la vie quotidienne

Patrice Guyot

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *21/03/2014* **par :**

Patrice Guyot

Caractérisation et reconnaissance de sons d'eau pour le suivi des activités de la vie quotidienne. Une approche fondée sur le signal, l'acoustique et la perception.

JURY

M. Georges Linares	Pr. Université d'Avignon	Président du Jury
M. Hervé Glotin	Pr. Université de Toulon	Rapporteur
Mme Jenny Benois-Pinot	Pr. Université de Bordeaux I	Examinatrice
M. Mathieu Lagrange	CR. Ecole Centrale Nantes	Examinateur
Mme Régine André-Obrecht	Pr. Université de Toulouse III	Directrice de thèse
M. Julien Pinquier	MCf. Université de Toulouse III	Co-directeur de thèse

---

École doctorale et spécialité :
MITT : Image, Information, Hypermedia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)
Directeur(s) de Thèse :
Régine André-Obrecht et *Julien Pinquier*
Rapporteurs :
Georges Linares et *Hervé Glotin*

Remerciements

Alors que je marchais dans les montagnes marocaines, en vacances après mon Master2, j'ai reçu un courriel de Régine André-Obrecht, me proposant de venir travailler en thèse sur le projet IMMED. Au delà de commentaires sur la réception de messages électroniques en toute occasion, je voulais remercier ma directrice de thèse pour cette attention qui m'a permis de me lancer dans ce travail de recherche. À de nombreuses reprises, Régine a su faire des retours très efficaces sur la manière de présenter et de valider mes approches.

Au quotidien, ce travail de thèse a été encadré par Julien Pinquier, que je tiens à remercier pour sa grande disponibilité, son enthousiasme et son implication. Ce travail à vocation interdisciplinaire n'aurait pas été possible sans l'importante autonomie que m'ont laissé mes encadrants pendant ces trois années, et je leur suis reconnaissant de la confiance qu'ils m'ont accordée.

Merci aux rapporteurs de cette thèse, Georges Linares et Hervé Bredin, qui, après avoir accepté de relire mon manuscrit et enduré de nombreuses heures de train, ont su formuler des retours constructifs, et installer un climat bienveillant lors de ma soutenance. Je remercie Mathieu Lagrange, qui, s'il a pris l'avion pour venir à ma soutenance, n'en fut pas moins un interlocuteur très pertinent et apprécié. Je remercie également Jenny Benois-Pinot, pour avoir accepté d'être examinatrice, et pour ses invitations régulières au LABRI de Bordeaux lors des réunions du projet IMMED.

J'en profite pour saluer chaleureusement les médecins, psychologues, chercheurs, doctorants et ingénieurs que j'ai rencontré dans le cadre de ce projet.

Je salue également tous les membres de l'équipe SAMOVA. J'ai une pensée particulière pour mes collègues de bureau durant cette thèse : Frédérique Gianni qui, alors que j'étais en début de thèse, a essayé de m'apprendre à lire un manuel, Maxime Le Coz pour son humour et ses chorégraphies, et Monsieur Lachachi qui m'a fait découvrir le sens de l'expression *l'heure des braves*.

Au cours de mes recherches, j'ai eu la chance de collaborer avec un doctorant catalan, Xavier Valero, que je remercie d'avoir eu la bonne idée d'effectuer une visite à Toulouse.

J'ai aussi profité de ce doctorat pour effectuer un séjour à l'IRCAM, qui a pu se concrétiser grâce à une bourse du GDR-ISIS. Au sein de l'équipe Perception et Design Sonores, j'ai été accueilli très amicalement par Patrick Susini, Nicolas Misdariis, ainsi qu' Olivier Houix avec qui j'ai réalisé la majeure partie de cette collaboration. Au risque de paraitre naïf, je dirai que ce séjour m'a appris à écouter les sons plutôt que les regarder. Je remercie tous les membres de l'équipe pour cette collaboration aussi enrichissante qu'agréable. Un grand merci également à toutes les personnes qui ont participé à mes expériences et qui ont subi l'écoute de mes sons de liquides sans rechigner (ou presque).

Merci à Pascal Gaillard et à l'équipe PETRA, voisins sympathiques et chaleureux, pour leur intérêt et pour leur investissement dans nos collaborations interdisciplinaires.

Je salue également les personnes que j'ai côtoyé durant mon parcours de recherche depuis mon Master1, qu'ils soient étudiants, enseignants et/ou chercheurs. Je tiens particulièrement à remercier mes collègues du parcours ATIAM, de l'Université de Montréal, les participants aux

JJCAAS, ceux rencontrés dans d'autres conférences, et tout ceux qui, par leur travail, leurs enseignements, ou tout simplement leur manière d'être, ont suscité chez moi l'envie d'évoluer dans le domaine de la recherche.

En dehors de ce contexte professionnel, je remercie mes proches pour leur présence réconfortante au cours de ces dernières années. Merci aux amis qui m'ont accompagné au fil de ce parcours, à mes collègues musiciens, et à ma famille. Je remercie particulièrement mes parents qui ne sont probablement pas complètement étrangers à la réalisation de ce travail, ni à son entreprise.

Merci à Aurélie qui m'a toujours encouragé quels que soient mes choix, et qui a infailliblement cru en mes capacités à porter et mener à bien ce projet.

Résumé

Avec le vieillissement de la population, le diagnostic et le traitement des démences telle que la maladie d'Alzheimer constituent des enjeux sociaux de grande importance. Le suivi des activités de la vie quotidienne du patient représente un point clé dans le diagnostic des démences. Dans ce contexte, le projet IMMED propose une utilisation innovante de la caméra portée pour le suivi à distance des activités effectuées. Nous avons ainsi travaillé sur la reconnaissance de sons produits par l'eau, qui permet d'inférer sur un certain nombre d'activités d'intérêt pour les médecins, dont les activités liées à l'alimentation, à l'entretien, ou à l'hygiène.

Si divers travaux ont déjà été effectués sur la reconnaissance des sons d'eau, ils sont difficilement adaptables aux enregistrements de la vie quotidienne, caractérisés par un recouvrement important de différentes sources sonores. Nous plaçons donc ce travail dans le cadre de l'analyse computationnelle de scènes sonores, qui pose depuis plusieurs années les bases théoriques de la reconnaissance de sources dans un mélange sonore.

Nous présentons dans cette thèse un système basé sur un nouveau descripteur audio, appelé couverture spectrale, qui permet de reconnaître les flux d'eau dans des signaux sonores issus d'environnements bruités. Des expériences effectuées sur plus de 7 heures de vidéo valident notre approche et permettent d'intégrer ce système au sein du projet IMMED. Une étape complémentaire de classification permet d'améliorer notablement les résultats. Néanmoins, nos systèmes sont limités par une certaine difficulté à caractériser, et donc à reconnaître, les sons d'eau.

Nous avons élargi notre analyse aux études acoustiques qui décrivent l'origine des sons d'eau. Selon ces analyses, les sons d'eau proviennent principalement de la vibration de bulles d'air dans l'eau. Les études théoriques et l'analyse de signaux réels ont permis de mettre au point une nouvelle approche de reconnaissance, fondée sur la détection fréquentielle de bulles d'air en vibration. Ce système permet de détecter des sons de liquide variés, mais se trouve limité par des flux d'eau trop complexes et bruités.

Au final, ce nouveau système, basé sur la vibration de bulles d'air, est complémentaire avec le système de reconnaissance de flux d'eau, mais ne peux s'y substituer. Pour comparer ce résultat avec le fonctionnement de l'écoute humaine, nous avons effectué une étude perceptive. Dans une expérience de catégorisation libre, effectuée sur un ensemble important de sons de liquide du quotidien, les participants sont amenés à effectuer des groupes de sons en fonction de leur similarité causale. Les analyses des résultats nous permettent d'identifier des catégories de sons produits par les liquides, qui mettent en évidence l'utilisation de différentes stratégies cognitives dans l'identification les sons d'eau et de liquide.

Une expérience finale effectuée sur les catégories obtenues souligne l'aspect nécessaire et suffisant de nos systèmes sur un corpus varié de sons d'eau du quotidien. Nos deux approches semblent donc pertinentes pour caractériser et reconnaître un ensemble important de sons produits par l'eau.

Abstract

The analysis of instrumental activities of daily life is an important tool in the early diagnosis of dementia such as Alzheimer. The IMMED project investigates tele-monitoring technologies to support doctors in the diagnostic and follow-up of the illnesses. The project aims to automatically produce indexes to facilitate the doctor's navigation throughout the individual video recordings. Water sound recognition is very useful to identify everyday activities (e.g. hygiene, household, cooking, etc.).

Classical methods of sound recognition, based on learning techniques, are ineffective in the context of the IMMED corpus, where data are very heterogeneous. Computational auditory scene analysis provides a theoretical framework for audio event detection in everyday life recordings. We review applications of single or multiple audio event detection in real life.

We propose a new system of water flow recognition, based on a new feature called spectral cover. Our system obtains good results on more than seven hours of videos, and thus is integrated to the IMMED framework. A second stage improves the system precision using Gammatone Cepstral Coefficients and Support Vector Machines. However, a perceptive study shows the difficulty to characterize water sounds by a unique definition.

To detect other water sounds than water flow, we used material provide by acoustics studies. A liquid sound comes mainly from harmonic vibrations resulting from the entrainment of air bubbles. We depicted an original system to recognize water sounds as group of air bubble sounds. This new system is able to detect a wide variety of water sounds, but cannot replace our water flow detection system.

Our two systems seem complementary to provide a robust recognition of different water sounds of daily living. A perceptive study aims to compare our two approaches with human perception. A free categorization task has been set up on various excerpts of liquid sounds. The framework of this experiment encourages causal similarity. Results show several classes of liquids sounds, which may reflect the cognitive categories.

In a final experiment performed on these categories, most of the sounds are detected by one of our two systems. This result emphasizes the necessary and sufficient aspect of our two approaches, which seem relevant to characterize and identify a large set of sounds produced by the water.

Celui qui, dans ses recherches scientifiques,
cherche à obtenir des applications pratiques immédiates,
peut être généralement assuré qu'il cherche en vain.

Hermann von Helmholtz

Table des matières

Introduction

1 Interdisciplinarité

1.1 Contexte universitaire

Le regroupement de plusieurs disciplines, au niveau de l'enseignement et de la recherche, est à la base de la construction des Universités. Ceci différencie en France les Universités des grandes écoles. Dans le monde, les universités les plus prestigieuses, par exemple l'Université Stanford, sont également bâties sur ce modèle de regroupement de disciplines, ce qui incite les étudiants issus de parcours différents à se rencontrer. L'interdisciplinarité crée l'ouverture et apporte un réseau.

Les vertus du cloisonnement des disciplines dans l'histoire des sciences, ne sont pourtant pas à démontrer. La spécialisation permet de mieux cerner les différents domaines et engendre régulièrement des avancées notoires. En France, le Conseil National des Université s'appuie ainsi sur une organisation hiérarchique des sections scientifiques, qui sont englobées dans des groupes. Il semble pourtant que de nombreuses thématiques de recherches actuelles se situent au croisement de plusieurs disciplines, tout comme l'a été l'écologie dans les années 30 [Tan35].

Au niveau individuel, l'ouverture des chercheurs à d'autres disciplines semble toujours enrichissante. Elle débouche parfois même sur des découvertes. Le météorologiste Alfred Wegener, par exemple, avait remarqué en regardant une mappemonde que l'Afrique de l'ouest et le Brésil s'ajustaient l'un à l'autre. En appuyant sa thèse sur des similitudes de faune et de flore sur les deux continents, il avait élaboré, en 1912, la théorie de la dérive des continents, qui fut longtemps refusée par les spécialistes du domaine [Mor90]. Si de tels exemples restent exceptionnels, il semble également que l'ouverture à d'autres champs permette pour le moins d'initier des idées, ou d'acquérir une vision plus globale d'une problématique.

Je conclurai le premier paragraphe de cette thèse effectuée à l'Université Paul Sabatier, par une citation de Blaise Pascal dont le triangle ornemente la station de métro de l'Université. Ainsi, selon le mathématicien, physicien, philosophe et théologien : « toutes choses étant causées et causantes, aidées et aidantes, médiates et immédiates, et toutes s'entretenant par un lien naturel et insensible qui lie les plus éloignées et les plus différentes, je tiens impossible de connaître les parties sans connaître le tout, non plus que de connaître le tout sans connaître particulièrement les parties [1] ».

1.2 Thématiques de recherche

En informatique, le traitement de données multimédia est un domaine qui se prête particulièrement bien à la rencontre des compétences. Ainsi, dans le domaine de la synthèse d'image, les

1. Blaise Pascal (1623-1662), Les pensées. Édition Lafuma, Paris, 1963.

chercheurs associent des connaissances d'optique, à du traitement du signal et des optimisations informatiques de calcul parallèle.

Dans l'équipe SAMOVA de l'IRIT, nous avons travaillé sur le traitement de données sonores. Un phénomène sonore peut être décrit de la manière suivante (voir figure 1). Les sources acoustiques en vibration créent une onde sonore qui se propage, par exemple dans l'air. Cette onde sonore peut être considérée sous la forme d'un signal analogique (tel qu'il est transmis par un haut-parleur ou un microphone) ou numérique (par exemple lorsqu'il est analysé de manière informatique). Elle peut également être entendue ou perçue par un ou plusieurs auditeurs.

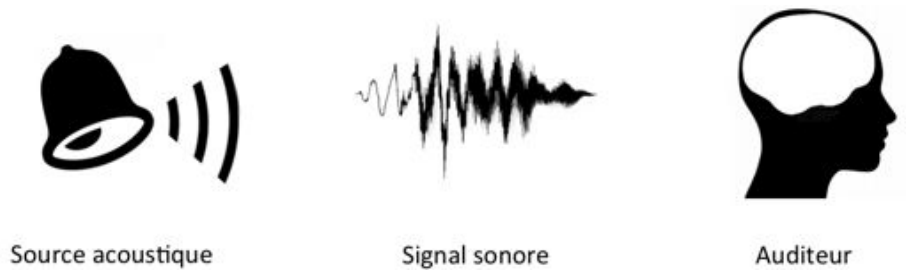


FIGURE 1 – Transmission d'un signal sonore.

Nous pouvons ainsi observer le phénomène sonore selon ces différents points de vue : la source acoustique, le signal sonore ou l'auditeur. A chacun de ces points de vues, peuvent être associés une ou plusieurs disciplines de recherche. De nombreux travaux se situent également au carrefour de ces disciplines.

Dans le domaine de l'analyse sonore, qui constitue notre cadre de recherche, l'objectif est ainsi d'obtenir à partir du signal des informations sémantiques de haut niveau sur les sources sonores. Par exemple, dans le domaine *Music Information Retrieval*, il s'agit de retrouver des concepts musicaux à partir du signal. Cette approche peut être évaluée par rapport à des auditeurs, qui, à partir du signal sonore également, sont amenés à créer une vérité terrain ou établir une référence.

Dans un autre domaine, la synthèse sonore, les recherches s'articulent entre le signal sonore synthétisé et sa perception. Toutefois, il est courant que les études s'appuient également sur des modèles de vibration des sources à reproduire, par exemple dans le cas de la synthèse par modèle physique.

Certaines études proposent enfin l'analyse d'un son particulier à travers les points de vue de la source physique en vibration, des propriétés du signal, et de la perception qu'en ont les auditeurs. L'une d'elles propose, par exemple, d'explorer les conséquences de la variations de la taille d'une plaque en vibration en terme d'acoustique, de signal produit et de perception [GM06]. Cette approche permet ainsi d'obtenir une vision assez large d'un phénomène sonore produit par un seul type de sources.

Il semble néanmoins pour l'instant que ces approches se soient focalisées sur les sons produits par des objets solides. Dans cette thèse nous allons investir ces trois points de vue différents pour considérer une problématique unique : reconnaitre les sons produits par l'eau et d'autres liquides.

2 Les sons d'eau

2.1 Les sons d'eau dans la nature : l'invisible orchestre des eaux

« Le premier son entendu ? Ce fut la caresse de eaux ». C'est ainsi que R. Murray Schaffer commence son fameux livre « Le paysage sonore » [Sch77]. De par l'origine marine des êtres vivant sur terre, les sons produits par l'eau figurent probablement parmi les premiers sons entendus. Depuis l'antiquité et de tout temps, ces sons ont inspiré poètes et écrivains qui ont versé des flots d'encre à produire des descriptions de paysages.

La mer, tout d'abord, dont le son des vagues n'est pas étranger au charme « je l'entends, elle bouge, elle respire [2] », ou à la frayeur qu'elle inspire « La création aveugle hurle, glapit, grince et beugle [3] ». Rares sont, de plus, les paysages qui soient aussi fortement associés aux sons qu'ils produisent que les paysages de mer [CBdL99]. Enfin, si la glace permet de conserver les paroles gelées chez Rabelais, qui évoque bien avant l'heure l'enregistrement sonore [4], « L'ambition de la mer » est également décrite comme « ce que la glace n'ose dire [5]».

Au delà des mers, les sons d'eau se retrouvent sur terre. Ils évoquent en général une image positive et relaxante [Bjö86]. « Le passant qui se tiendrait immobile dans cette partie de la campagne, par une nuit paisible, entendrait de singulières symphonies jouées par l'invisible orchestre des eaux [6] ». Le son de ruisseau est ici décrit comme un « doux murmure », tandis que parfois le bruit d'une rivière peut être « semblables aux détonations lointaines des fusils [7] ».

Dans la nature, les sons d'eau nous parviennent également par la pluie. « Il n'est pas deux gouttes de pluies qui sonnent de la même manière (...) Comment la pluie persane pourrait elle ressembler à celle qui tombe sur les Açores ? » [Sch77].

2.2 Variabilité

Nous voyons à travers ces exemples littéraires que les sons produits par l'eau possèdent une place importante dans notre perception sonore du monde extérieur. Ces descriptions s'appuient sur un vocabulaire varié.

Cette variabilité de vocabulaire se retrouve dans les verbes, utilisés pour décrire le mouvement ou les actions liées aux liquides : *ruisseler, déferler, couler, arroser, asperger, baigner, barboter, bouillir, écoper, égoutter, épancher, goutter, jaillir, puiser, patauger, patouiller, verser, renverser.* Chacune de ces actions va produire un son. De plus, malgré le nombre important de termes utilisés pour décrire ces actions, il semble envisageable d'associer un son différent à chacun de ces termes.

Les manifestations de l'eau peuvent également être décrites par un vocabulaire important, par exemple : *mer, lac, cascade, rivière, torrent, fleuve, marais, lagune, flaque, filet.* Ces différentes manifestations de l'eau peuvent également produire des sons très spécifiques.

Certains sons d'eau peuvent enfin être décrits directement par des termes particuliers : *clapotis, glouglou, gargouillis.*

La quantité de termes utilisés pour décrire l'élément liquide et les sons associés confirme d'une part l'importance de cet élément dans notre environnement. De plus elle suggère une variété importante de sons différents, pouvant être associés à un paysage ou une action précise.

2. Jean-Marie Gustave Le Clézio, Le Chercheur d'or.
3. Victor Hugo. « Océan » dans La Légende des siècles.
4. François Rabelais. Gargantua.
5. Jean-Claude Izzo, Loin de tous rivages
6. Thomas Hardy, Le maire de Casterbridge
7. J. Fenimore Cooper, The Pathfinder, New York, 1863.

Certains de ces termes, comme par exemple *bouillir*, n'évoquent pas des paysages extérieurs, mais plutôt des actions ou des activités effectuées au quotidien dans notre domicile. Le site internet *soundfishing*[8], permettant de télécharger des sons, propose 176 bruitages dans la catégorie « bruitage d'eau ». Dans ce site, ces sons sont classés selon les thèmes *bruit de WC, bulles, chute d'eau, cours d'eau, cuisine, fontaine, goutte d'eau, moulin écluse, plongée sous-marine, plongeon, salle de bain*. Parmi ces différentes catégories, trois correspondent à des lieux précis du domicile : la cuisine, la salle de bain et les toilettes.

2.3 Les sons d'eau dans les activités humaines

Si les sons d'eau peuvent facilement évoquer des paysages, ils sont également très présents dans notre quotidien. En effet, boire est l'un des besoins primaires de nombreux êtres vivants. La proximité de l'eau a donc été une contrainte importante pour la sédentarisation des populations. La mise en place de l'eau courante, puis de l'eau chaude dans les foyers a amélioré successivement les conditions de confort.

Aujourd'hui, l'eau et les liquides interviennent premièrement dans les besoins d'alimentation. Ces liquides sont ainsi utilisés en tant que boisson, ou pour cuisiner. De ces activités naissent plusieurs types de bruits, par exemple les sons de remplissage et du versement de divers types de contenants. Nous pouvons ajouter à ces exemples les sons d'un plat qui mijote, de l'eau qui bout, ou de certaines cafetières. L'eau est de plus régulièrement associée à l'hygiène et permet de se laver. Les lavabos, douches, baignoires et toilettes produisent différents types de son. Le jet du pommeau de douche peut ainsi difficilement être confondu avec celui d'un robinet de baignoire. Les liquides dont l'eau sont également utilisés pour nettoyer. Leur utilisation dans les activités de ménage est également audible : le remplissage d'un évier avant de faire la vaisselle, les mouvements d'eau dans le bac, l'essorage de l'éponge, ou l'utilisation d'une serpillère. Enfin, l'eau peut être aussi associée à certaines activités relatives aux loisirs, comme la piscine ou le jardinage.

L'eau et les liquides sont ainsi utilisés dans des activités variées de la vie courante. La plupart de ces activités produisent des sons. Certains de ces sons, comme la chasse d'eau, sont bien spécifiques. Si certains autres sons ne permettent pas d'identifier de façon sûre une activité, ils peuvent en revanche nous permettre d'obtenir des indices sur l'activité effectuée. Les sons de liquide peuvent ainsi nous permettre d'inférer sur les activités effectuées et peuvent être utilisés dans des applications médicales.

3 Contexte de travail

L'observation des activités du quotidien devient en effet depuis quelques années un outil incontournable pour le diagnostic et le suivi des démences, telles que la maladie d'Alzheimer. Le suivi de ces activités permet en effet au médecin d'effectuer un diagnostic précoce, de suivre l'évolution de la maladie et de statuer sur la capacité du patient à vivre de manière autonome.

Nous avons travaillé dans cette thèse sur le projet IMMED. Ce projet, à vocation interdisciplinaire, rassemble des médecins et des spécialistes du traitement audio et vidéo. Le but de ce projet est de permettre aux médecins d'évaluer la capacité des patients à réaliser une série d'activités de la vie quotidienne au sein de leur lieu d'habitation. L'observation du patient dans son propre lieu d'habitation conduit en effet à une évaluation plus objective de ses capacités à réaliser les activités de la vie quotidienne. L'utilisation de la vidéo permet au médecin d'observer

8. http://www.sound-fishing.net/bruitages_eau.html, consulté le 23 octobre 2013.

à distance ces différentes activités. Au final, le visionnage de ces vidéos s'ajoute aux autres outils de diagnostic du médecin (entretien avec les patients et les proches, imagerie cérébrale, etc).

Un des objectifs du projet est la segmentation automatique de la vidéo en activités réalisées par le patient afin de faciliter son visionnage par le médecin. Nous avons travaillé à l'IRIT sur l'analyse sonore permettant de trouver des informations sur l'activité des patients. Comme chaque vidéo est tournée dans des domiciles différents et dans des conditions relativement bruitées, nous nous sommes rapidement tournés vers l'analyse des sons d'eau, qui sont présents dans de nombreuses activités d'intérêt, et qui ont l'avantage d'être relativement constants selon les différents domiciles des patients.

Nous nous plaçons donc dans la thématique de l'analyse automatique des sons environnementaux. Ces sons environnementaux, couramment appelés « bruits », sont généralement opposés à la parole et à la musique. Dans le domaine du traitement du signal audio, l'analyse de ses sons est très minoritaire par rapport aux recherches sur la parole ou la musique. Ce constat est étayé par le fait que la plupart des travaux de recherches effectués dans ces derniers domaines s'appuient sur des corpus enregistrés en studio. Ce type de corpus est en général composé exclusivement de parole et de musique.

4 Contributions

La reconnaissance robuste des sons d'eau constitue notre axe de travail. Pour des raisons pratiques liées à notre application, nous nous sommes focalisés sur les sons de liquides issus des activités de la vie quotidienne. Le suivi d'activité à partir de l'analyse des sons d'eau a déjà été utilisé dans différentes études pour des applications médicales. D'autres applications existent également dans le domaine de l'écologie. Au niveau du traitement du signal, la plupart de ces études utilisent des outils très génériques, et ne se basent pas sur une analyse du son à reconnaître. Pour résoudre notre problématique nous avons analysé la faculté des descripteurs audio à modéliser les sons de liquide. Cette analyse nous a conduits à la proposition d'une nouvelle méthode basée sur un descripteur original.

Notre approche issue de l'observation du signal s'est pourtant trouvée limitée par la variabilité des sons d'eau. Ainsi, comme les sons d'eau dans la nature peuvent être associés à la mer, de ruisseau, de rivière et de pluie, les sons d'eau produits par les activités humaines peuvent prendre un certain nombre d'aspects différents. Ces sons, acoustiquement très différents, proviennent pourtant tous d'un phénomène physique, connu depuis près d'un siècle. Inspirés par les études de synthèse sonore de sons d'eau, nous avons étudié les modèles physiques à la base de la production de son. Cette étude a débouché sur un nouveau système permettant de détecter un autre groupe de sons d'eau.

Au final, les deux approches choisies se sont montrées complémentaires et sont toutes deux sont nécessaires à la détection de l'ensemble des sons d'eau de notre corpus. Pour expliquer ce phénomène et justifier de l'aspect nécessaire et suffisant de ces deux approches, nous nous sommes alors tournés vers la perception sonore. Nous avons donc effectué une série d'expériences avec l'équipe « Perception et Design Sonores » de l'IRCAM. Ces expériences mettent en évidence l'utilisation de processus cognitifs différents dans l'écoute et la reconnaissance d'un ensemble varié de sons d'eau du quotidien. Ces différents processus cognitifs peuvent être comparés avec nos deux approches de reconnaissance automatique.

5 Plan de la thèse

La première partie de cette thèse présente notre cadre de travail. Le premier chapitre est ainsi dédié aux applications médicales des sciences et technologies de l'information. Il présentera de plus le projet IMMED et ses problématiques qui nous ont orientées vers l'étude des sons d'eau. Dans la deuxième chapitre, nous proposons un état de l'art des techniques de traitement du signal audio, sous l'angle de la reconnaissance automatique des sons environnementaux.

La deuxième partie de cette thèse décrit de manière précise nos contributions. Chacun des trois chapitres est dédié à l'un des points de vue du phénomène sonore: signal, vibratoire, et perceptif. Le chapitre 3 de la thèse présente ainsi nos méthodes de détection de flux d'eau. Ces méthodes sont évaluées sur des enregistrements sonores de la vie quotidienne issus du projet IMMED. Le chapitre 4 propose une nouvelle approche basée sur les modèles acoustiques de sons de liquide. Une étude perceptive présentée dans le chapitre 5 permet de comparer l'audition humaine de ce type de sons à nos approches automatiques.

Chapitre 1

Reconnaissance automatique d'activités de la vie quotidienne pour l'aide au diagnostic

Résumé du chapitre : Le suivi à distance des activités de la vie quotidienne du patient constitue une étape importante dans le diagnostic et le traitement de démences comme la maladie d'Alzheimer. Les Sciences et Technologies de l'Information et de la Communication trouvent une application médicale dans le suivi à distance des patients par l'intermédiaire de capteurs fixes ou portables. Le projet IMMED propose ainsi une utilisation innovante de la caméra portée pour le suivi d'activités de la vie quotidienne. Le visionnage des vidéos par les médecins s'appuie sur un modèle de reconnaissance automatique d'activités. La reconnaissance d'évènements sonores spécifiques, tels que les sons d'eau, constitue un indice utile pour la segmentation en activités. Elle semble de plus réalisable malgré des conditions d'enregistrements hétérogènes.

1.1 Contexte général

Avec l'augmentation de l'espérance de vie, le diagnostic et le traitement des démences sont devenus des enjeux sociétaux et économiques importants dans les pays riches. En France, d'après la fondation plan Alzheimer [Alz13], le nombre de démences a été évalué à 1,2% de la population en 2010, chiffre qui devrait être multiplié par deux d'ici 2050. La maladie d'Alzheimer est la plus fréquente des démences du sujet âgé, avec environ 70% des cas. Elle affecte de manière importante les proches du patient sur qui repose le plus souvent la lourde charge humaine, affective et financière.

Les démences telles que la maladie d'Alzheimer sont habituellement diagnostiquées par l'intermédiaire d'indices sur des changements pathologiques dans la vie de tous les jours des patients. Pour évaluer ces changements, les médecins utilisent des examens neurologiques, des techniques d'image cérébrale, et principalement des tests neuropsychologiques [MDF$^+$84]. Actuellement, le processus de diagnostic de ces démences ne permet pas d'identifier à temps toutes les personnes souffrantes dans la population. Des études à grandes échelles, comme la campagne PAQUID qui a impliqué près de 4000 patients du sud-ouest de la France de 1998 à nos jours montrent en effet que des signes précoces de la maladie peuvent apparaître près de 10 ans avant son diagnostic clinique par les tests neuropsychologiques [PHA$^+$08]. Un diagnostic précoce de la démence permet une meilleure prise en charge du patient et de ses proches, et ouvre la voie à de nouvelles approches thérapeutiques susceptibles de freiner ou d'arrêter la progression de la maladie.

La déficience dans la réalisation des activités de la vie quotidienne est un indice important du développement futur de la maladie d'Alzheimer. L'évaluation de cette déficience est difficile, et dépend généralement d'outils subjectifs et de compétences à analyser clairement la situation. Elle se heurte au déni du patient et de ses proches, ou à l'anosognosie. Concrètement, les équipes médicales estiment le plus souvent des aptitudes énoncées par le patient lui-même ou par ses proches. Le développement de nouveaux protocoles et outils pour évaluer la capacité du patient à réaliser certaines activités de la vie quotidienne est donc un enjeu important pour le diagnostic précoce des démences.

Les récents progrès dans les Sciences et Technologies de l'Information et de la Communication (STIC) conduisent de nos jours à de nombreuses applications novatrices, notamment dans le domaine de l'aide aux personnes âgées. L'exemple de ces applications le plus couramment utilisé est peut-être le détecteur de chute (voir [HAAH10] pour un état de l'art), qui permet de prévenir rapidement les secours quand une personne âgée chute, et garantit ainsi une relative sécurité du patient en situation d'autonomie. D'autres applications utilisent des capteurs pour mesurer et suivre l'évolution des paramètres physiologiques du patient (voir [NFN+09] pour une présentation de ces capteurs). Nous allons voir que d'autres études se focalisent sur la reconnaissance automatique d'activités de la vie quotidienne, dont l'un des objectifs est le maintien des patients à domicile (voir [Ram10] pour plus de détails sur les applications de maintien à domicile).

Dans ce chapitre, nous allons nous intéresser à la reconnaissance d'activités à domicile dans l'objectif de l'aide au diagnostic précoce de la maladie d'Alzheimer. La section 1.2 présente un rapide inventaire des études de reconnaissance automatique d'activités dans le domaine médical. Les sections 1.3 et 1.4 décrivent respectivement le projet ANR Blanc IMMED [9] qui a été le cadre initial de cette thèse, et le corpus réalisé dans ce projet. La section 1.5 conclut ce chapitre par les problématiques liées à la reconnaissance automatique d'évènements sonores dans le corpus IMMED qui nous ont amenés à nous focaliser sur les activités liées à l'utilisation de l'eau.

1.2 Les sciences et technologies de l'information et de la communication pour la reconnaissance d'activités

Nous allons passer en revue les principales utilisations des STIC dans le domaine de la reconnaissance et du suivi à distance des activités du patient. Nous nous limitons aux technologies liées à la détection de présence, de mouvements, ainsi que la capture d'images et de sons. Ainsi, le suivi à distance des paramètres physiologiques du patient n'est pas abordé. Les applications liées spécifiquement à la reconnaissance de sons sont détaillées dans le chapitre 2 et les travaux utilisant la reconnaissance de sons d'eau sont présentés dans le chapitre 3.

1.2.1 Différents types d'activités

Définition de l'activité

Dans le langage courant, le terme *activité* se réfère aux actions effectuées par un individu :
- *Action de quelqu'un* (Petit Larousse en ligne),
- *Ensemble des actes coordonnés et des travaux de l'être humain* (Petit Robert).

Ce terme est aussi utilisé dans la littérature liée aux STIC à propos d'une posture (*debout, couché, assis*) ou d'un type de déplacement (*marcher, courir*) [HAAH10].

9. http://www.immed.labri.fr

Dans le domaine médical, l'expression « Activités de la Vie Quotidienne » (AVQ) est également fréquemment utilisé [Fri13]. Les AVQ se divisent en activités instrumentales de la vie quotidienne (AIVQ) et soins personnels ou AVQ de base (AVQB). Les AIVQ peuvent être définies comme « relatives à la capacité de la personne de se débrouiller dans son environnement par rapport à des tâches adaptées comme faire des achats, cuisiner, faire le ménage, faire la lessive, utiliser un transport, gérer son argent, gérer ses médicaments et utiliser le téléphone » [Kat83]. Les AVQB étant « restreintes aux activités impliquant la mobilité fonctionnelle (marche, mobilité en fauteuil roulant, mobilité dans le lit et transferts) et les soins personnels (alimentation, hygiène, élimination, bain et habillage) » [RL08]. Le terme AVQ est également couramment utilisé dans le domaine des STIC pour se référer à ces deux ensembles d'activités (AIVQ et AVQB).

La mobilité, une activité primaire

Une des activités, la plus primaire et la plus facile à suivre à distance par l'intermédiaire des STIC, semble être la mobilité fonctionnelle du patient. Son étude apporte un indice important sur la santé du patient : « on constate de plus en plus que l'immobilité n'existe vraiment que dans la mort... » [10].

La mobilité constitue ainsi une des actions primaires que le patient peut effectuer, mais également la base nécessaire à la réalisation d'actions plus complexes : « une mobilité importante augmente l'endurance et la force musculaire, et peut améliorer le bien-être psychologique et la qualité de vie en augmentant les capacités des personnes à effectuer un large panel d'activités de la vie quotidienne » [Ser96].

De plus, l'absence de mobilité, éventuellement liée à la connaissance de la position du patient, peut permettre de détecter une situation dangereuse. Par exemple, dans certaines applications, le fait de savoir qu'une personne âgée reste dans le salon en position allongée va déclencher une alerte. Les méthodes automatiques de suivi de la mobilité ont été répertoriées dans [SCB+06] ainsi que dans [dBHU+08].

Activités complexes

Les activités plus complexes, dont les activités instrumentales mais aussi certaines AVQ de base liées à l'hygiène ou à l'alimentation sont beaucoup plus difficiles à reconnaître de manière automatique. Par exemple, l'activité *prendre ses médicaments* nécessite une compréhension des gestes du patient et de son environnement. La reconnaissance d'activités complexes constitue néanmoins un défi majeur pour le suivi à distance des personnes âgées. Par exemple, la reconnaissance de l'activité *prendre des médicaments* peut être d'importance vitale sur la santé du patient.

D'autres activités, comme celles liées à l'hygiène, peuvent causer une dégradation générale de l'état de santé du patient si elles ne sont pas effectuées. D'une part, il est utile de savoir si une personne atteinte de démence se lave les mains régulièrement [TSGM10] (cette dernière étude sera présentée en détail dans le chapitre 3). D'autre part, le suivi d'activités comme *faire la cuisine* ou *faire le ménage* permet d'appréhender l'autonomie du patient au sein de son lieu de vie. Enfin, certaines activités permettent de constater l'intégration du patient dans son environnement social, comme le fait d'utiliser un téléphone, les visites de proches, la lecture ou l'utilisation de médias.

Dans [PFP+04], les auteurs proposent de détecter automatiquement 14 AVQ : *apparence personnelle, hygiène orale, faire sa toilette, laver, nettoyer, utiliser l'électroménager, utiliser le*

10. Harvey, J. C. (1988). *Les demi-civilisés*. Montréal: Presses de l'Université de Montréal.

four, faire la lessive, préparer un plat, préparer une boisson, utiliser le téléphone, loisirs (lecture, télévision, etc.), s'occuper d'un enfant, prendre ses médicaments. Cette détection se fait grâce à des capteurs placés sur les objets de la maison, et à l'utilisation d'un gant spécial permettant de les détecter (voir partie 1.2.3). Cette étude propose ainsi un vaste panel d'activités à reconnaître, ce qui est très utile d'un point de vue médical. D'un point de vue pratique par contre, l'aménagement du domicile avec les capteurs et l'utilisation du gant peut constituer un protocole contraignant. Ce type d'application semble contraint par le compromis entre la qualité de la reconnaissance et la lourdeur du dispositif expérimental.

Vie privée

Si le suivi continu d'un grand nombre d'activités améliorerait notablement le diagnostic du médecin et pourrait rassurer les proches du patient, l'enregistrement systématique de la vie du patient se heurte déontologiquement au respect de la vie privée. Le suivi précis d'activités complexes est donc difficilement envisageable de manière continue sur la journée.

Dans le domaine de la reconnaissance d'activités, le choix du type de capteur est ainsi un paramètre important pour le respect de la vie privée. Ainsi, pour une application de détection de chute, un accéléromètre ne dévoilera pas d'informations intimes alors qu'une caméra fixe utilisée en continu sera perçue comme très intrusive.

Les parties suivantes détaillent les différents types de capteurs généralement utilisés dans les applications de reconnaissance automatique d'activités.

1.2.2 Capteurs fixes

Dans les applications de suivi d'activités, les capteurs fixes sont habituellement installés de manière durable dans des lieux. Ces lieux sont couramment appelés « maisons intelligentes ».

Maisons intelligentes

Le premier projet de suivi d'activités dans une maison intelligente a été présenté en 1994 [CHEI94]. Dans cette étude, des capteurs de présence et de température communiquent avec un centre de calcul à distance via le réseau téléphonique, et un capteur sonore permet de préciser le type d'activités.

Aujourd'hui, il semblerait que le terme *maison intelligente* puisse correspondre à deux catégories de projets. Les capteurs peuvent premièrement être installés dans des lieux spécifiques, par exemple à l'hôpital, où différents patients se rendent de manière ponctuelle. Le système PROSAFE (Surveillance automatisée et non intrusive de personnes âgées et/ou dépendantes pour une aide à l'autonomie) propose d'identifier de manière automatique les activités quotidiennes de personnes suivies à leur domicile par une période d'apprentissage de 30 jours, puis de détecter toute anomalie de comportement comme une chute, une fugue, ou une agitation nocturne. L'extension de ce modèle à des domiciles personnels a été considérée dans [BCEG07], mais nécessite des moyens techniques très importants.

Ces maisons intelligentes servent également de test pour un équipement futur des maisons des patients, afin de permettre un maintien à domicile. Dans ce dernier cas, la maison du patient est directement équipée de capteurs. Cette deuxième application est soumise à la difficulté d'équiper des domiciles personnels ainsi qu'au coût d'une telle opération. De nombreuses études scientifiques décrivent des prototypes d'implémentation à domicile ; toutefois l'utilisation commerciale de dispositifs fixes dans les maisons individuelles semble encore limitée.

Dans la littérature scientifique, ces deux types d'applications sont couramment réunies sous le terme *maison intelligente* : voir [CEEC08] pour une synthèse sur les maisons intelligentes et [Ram10] pour les applications orientées vers le maintien à domicile.

Capteurs de position

Les capteurs de position sont utilisés pour détecter la présence ou l'absence d'individus dans une zone spécifique. Les capteurs infrarouges détectent le mouvement du corps humain par la mesure du rayonnement infrarouge, qui correspond à la chaleur émise par le corps humain. Ils fournissent une indication de changement d'occupation d'une pièce : absence ou présence. Ils peuvent être ainsi utilisés pour suivre les déplacements du patient et avoir une idée de la distance parcourue [CHRC95]. Ils ne permettent pas en revanche de connaître le taux d'occupation d'un local ou le nombre d'occupants. Un capteur infrarouge est représenté sur la figure 1.1.

FIGURE 1.1 – Détecteur passif infrarouge.

Les capteurs de position semblent assez adaptés pour suivre la mobilité du patient de façon non-intrusive. Il est néanmoins nécessaire de les installer préalablement dans chaque pièce du lieu de vie. Au-delà des déplacements, il est difficile de déduire précisément les activités réalisées par l'intermédiaire de ces capteurs.

Microphones

Les microphones fixes permettent d'enregistrer les scènes sonores se déroulant dans les lieux de vie des patients. Si l'écoute de ces enregistrements peut fournir de nombreuses informations sur les activités du patient, ces capteurs audio sont concrètement peu utilisés dans les projets de maisons intelligentes. Leur utilisation automatique nécessite en effet des algorithmes complexes de traitement du signal audio pour détecter les activités sans fausses alarmes [SCB+06]. Ils ont cependant l'avantage d'être moins intrusifs que les caméras, peu coûteux, et de dimension réduite (voir figure 1.2).

L'équipe du laboratoire CLIPS-IMAG de Grenoble a étudié l'utilisation de ces capteurs pour détecter des situations de détresse. Le traitement des signaux acoustiques enregistrés par plusieurs microphones disposés dans un appartement [CI01, VIB+03, ICV+06, AVRP13] permet de localiser une personne, ou de détecter les cris et les appels au secours.

Les microphones peuvent également permettre de détecter les chutes [PLSR08]. Ils sont aussi utilisés pour détecter des sons spécifiques (*évier, téléphone, douche, porte*) pour le suivi d'activités à distance [LAC07]. Nous nous focaliserons sur cette approche de détection de sons spécifiques dans cette thèse. Les algorithmes de traitement du signal utilisés et les résultats dans ce type de projet seront détaillés dans le chapitre 2.

FIGURE 1.2 – Microphone pour le suivi à distance.

Vidéo

Les caméras sont fréquemment utilisées dans les maisons intelligentes. L'amélioration rapide des technologies conduit à des dispositifs aussi miniaturisés que performants (figure 1.3). L'utilisation de la vidéo pour reconnaître automatiquement les activités du patient constitue un problème complexe qui est régulièrement abordé dans la littérature scientifique. Une approche courante consiste par exemple à estimer l'arrière-plan afin de ne conserver que les personnes ou les objets en mouvement. Ce type d'algorithme peut permettre la détection d'une chute lorsque la personne reste allongée trop longtemps et ainsi déclencher une alarme [FAP08].

Cette approche peut aussi nous amener à la reconnaissance d'activités plus précises [ZCC+08]. Dans cette dernière étude, la vitesse et la position du patient sont utilisées pour identifier les principales phases de la vie quotidienne, telles que *marcher, s'asseoir, être dans la salle de bain, préparer à manger, s'asseoir à table.* L'identification d'actions plus précises semble toutefois difficile, et si la reconnaissance d'activités précises, telles que *faire la cuisine* ou *se brosser les dents*, est évoquée dans ce projet, la mise en place d'une expérience en condition réelle d'utilisation n'a pas été présentée.

FIGURE 1.3 – Caméra pour le suivi à distance.

Systèmes hybrides

Les approches hybrides qui utilisent différents types de capteurs sont très utilisées dans les maisons intelligentes. Un lieu de ce type a ainsi été proposé par l'équipe CHU-PULSAR de Nice

pour la reconnaissance d'activités [ZBT+09]. Cette maison utilise des caméras fixes, des capteurs de présence, mais également des capteurs de pression sur les chaises et des capteurs permettant de suivre la consommation d'eau.

Dans ce type de système hybride, des algorithmes de haut niveau sont nécessaires pour fusionner l'ensemble des données des capteurs afin de reconnaître les activités. Certaines applications utilisent par exemple la logique floue [MIBD09].

Conclusion sur les capteurs fixes

Les technologies disponibles aujourd'hui permettent l'équipement d'une maison par des capteurs infrarouges de position, des microphones et des caméras vidéo. La miniaturisation de ces technologies rend leur utilisation pratique acceptable, bien que coûteuse, dans le cadre quotidien. Toutefois, elle se heurte, selon les capteurs utilisés, à la problématique du respect de la vie privée.

Les algorithmes de traitement de l'information permettent le suivi de la position et des déplacements du patient, ce qui constitue une première approche vers le suivi à distance d'activités. Le suivi plus précis d'activités reste un problème ouvert et semble aujourd'hui difficile sans un grand nombre de capteurs.

1.2.3 Capteurs portables

Les systèmes portables semblent prometteurs pour le suivi à distance de l'activité des personnes âgées. Ils ont le grand avantage de ne pas nécessiter d'installation particulière au domicile, et d'être éventuellement utilisables dans plusieurs lieux ou à l'extérieur. Par contre, ils doivent être portés par le patient, parfois quotidiennement et de manière continue. Leur utilisation régulière peut ainsi constituer une gêne.

Accéléromètres et gyroscopes

Les accéléromètres permettent de mesurer une accélération linéaire, alors que les gyroscopes permettent de détecter une rotation. Ces capteurs sont couramment utilisés dans la détection de chute (figure 1.4). Ils ont aussi, grâce à leur petite taille, l'avantage de pouvoir être portés facilement par le patient, ou embarqués dans d'autres objets comme les téléphones portables, ce qui permet d'utiliser facilement ces derniers comme détecteurs de chute [ZWLH06].

FIGURE 1.4 – Détecteur de chute avec accéléromètre.

Ce type de capteur semble donc très efficace pour la détection de chute, qui est une cause principale d'accident chez les personnes âgées. Des algorithmes adaptés peuvent diminuer significativement les fausses alarmes, par exemple en différenciant plusieurs types de chutes de mouvements de la vie quotidienne [WCL+08]. L'utilisation d'accéléromètre pour la reconnaissance d'activités semble difficile et n'a pas encore été abordée à notre connaissance.

Capteurs RFID

Les capteurs RFID (Radio Frequency IDentification) peuvent être collés ou incorporés dans des objets. Ces radio-étiquettes comprennent une antenne associée à une puce électronique qui leur permet de recevoir et de répondre aux requêtes radio émises depuis l'émetteur-récepteur.

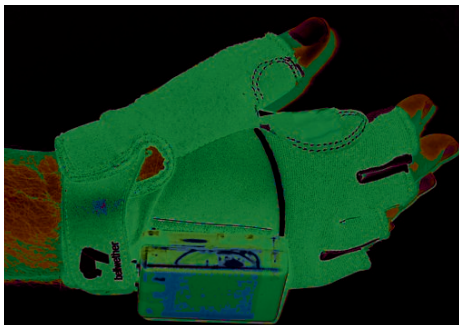


FIGURE 1.5 – Gant RFID.

Dans [PFP+04], les auteurs utilisent cette technologie en utilisant des radio-étiquettes sur les objets de la maison, alors que le patient utilise un gant détecteur (voir figure 1.5). Si cette technologie, peut sembler contraignante de nos jours, la rapide miniaturisation des équipements pourrait la rendre facilement opérationnelle dans les années futures.

Combinaisons de capteurs

De nombreuses recherches ont été effectuées grâce à des combinaisons de différents types de capteurs, qu'ils soient fixes ou portables. Par exemple, les capteurs RFID ont été combinés à des accéléromètres pour améliorer les résultats de différenciation des activités instrumentales des déplacements quotidiens [BUS+11].

Caméra embarquée

L'utilisation d'une caméra portée pour des applications médicales est plutôt originale. Un dispositif connaît néanmoins un succès important dans le domaine des troubles de la mémoire : le système SENSECAM qui enregistre des images de la vie quotidienne du porteur (voir figure 1.6). Le dispositif SENSECAM prend automatiquement des photos en fonction du mouvement et des changements d'image. Il garde une trace du quotidien des porteurs, et leur permet d'effectuer un travail clinique pouvant diminuer les troubles de la mémoire [HWB+06].

Une caméra portée a également été utilisée dans le projet WEARCAM (figure 1.6). Ce dispositif permet de suivre le regard du porteur. Il est utilisé sur de très jeunes enfants pour déceler des troubles de l'attention, et contribue à un diagnostic précoce de l'autisme [PNB+07].

Très récemment, une étude a été effectuée sur la détection de chute par une caméra portée [OMCV13]. Ce type d'étude confirme l'intérêt grandissant pour ces dispositifs vidéo portables. Ils ont l'avantage de pouvoir être utilisés partout sans installation de capteurs, et respectent la vie privée du patient qui n'est pas directement filmée. De plus, ils permettent l'utilisation d'algorithmes de reconnaissance audio et vidéo dont l'efficacité s'améliore continuellement.


FIGURE 1.6 – Systèmes d'acquisition SENSECAM (à gauche) et WEARCAM (à droite).

1.3 Le projet IMMED

1.3.1 Intérêt médical

Le projet ANR IMMED (Indexation de données MultiMédia Embarquées pour le diagnostic et le suivi des traitements des Démences) a été créé dans le domaine du diagnostic et du suivi des démences dans l'objectif de développer et de valider des technologies reposant sur la caméra portée [MDW+10].

Il s'agit de réduire le « manque d'objectivité » dans l'analyse des déficiences du patient à réaliser certaines activités de la vie quotidienne (cf. partie 1.1). Pour cela, il est nécessaire d'évaluer les éventuelles erreurs des patients au sein d'un environnement réel et habituel. Le projet propose ainsi aux médecins d'estimer les aptitudes du patient à exécuter les activités de la vie quotidienne au sein de son domicile.

Les activités d'intérêt, dont une liste est présentée en annexe 1, permettent d'évaluer les compétences praxiques du patient. Ces activités sont classées selon les catégories suivantes :

- entretien corporel et hygiène,
- alimentation,
- entretien du domicile,
- distractions,
- relations sociales.

L'idée du projet repose sur l'utilisation d'une caméra vidéo portative afin d'enregistrer les activités du patient à son domicile, et de permettre aux médecins de regarder ces vidéos directement sur leur lieu de travail. La visualisation de ces vidéos est ajoutée aux autres outils utilisés dans le diagnostic de démences : les tests, les entretiens avec le patient et les proches, etc.

1.3.2 Indexation automatique

L'acquisition vidéo au domicile du patient fournit des séquences assez longues. La durée de ces séquences varie de quelques dizaines de minutes à plus d'une heure, ce qui est trop long pour être visualisé entièrement par un spécialiste. De plus, les scènes filmées n'ont d'intérêt que si l'autonomie du patient peut être évaluée. La navigation dans les vidéos effectuées est un point critique pour la viabilité concrète du projet.

Le projet IMMED propose d'indexer les vidéos de manière automatique afin de les rendre utilisables concrètement par les médecins. L'annotation est effectuée selon les activités d'intérêt (voir annexe 1). Une interface de navigation utilise la segmentation automatique en activités pour proposer au médecin de regarder l'activité de son choix. Cette interface permet au spécialiste de visualiser directement les scènes où l'autonomie du patient peut être évaluée, ou de regarder la vidéo de manière séquentielle (voir figure 1.7).

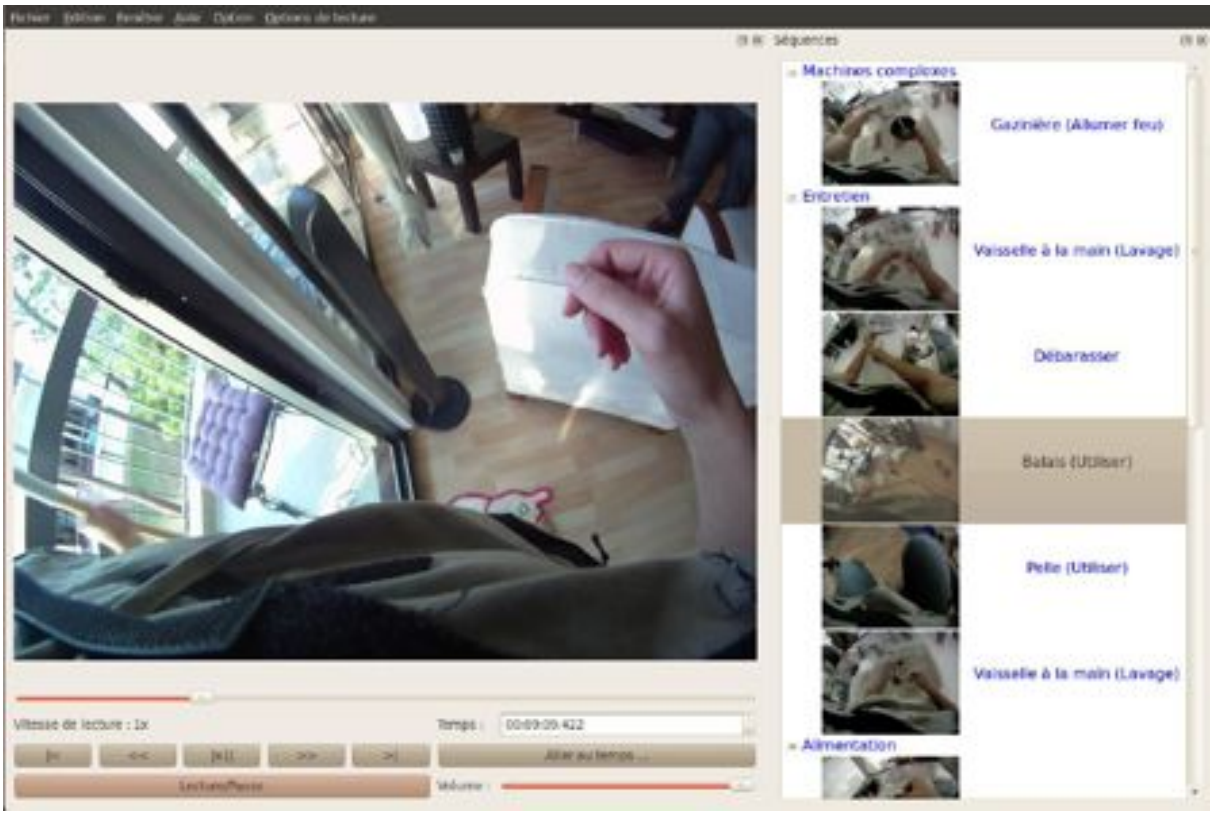


FIGURE 1.7 – Interface de visualisation des activités.

1.3.3 Scénario d'usage du projet

Dans un scénario typique d'utilisation (voir figure 1.8), le médecin ordonne une visite au domicile du patient. Un assistant médical se rend alors chez le patient muni du dispositif portable d'acquisition audio-vidéo. Il aide ensuite le patient à revêtir le dispositif et lance l'acquisition de la vidéo. Le dialogue avec le patient amène ce dernier à exécuter un ensemble d'activités de la vie quotidienne ciblées par le médecin.

Ces activités sont enregistrées par le dispositif d'acquisition. Depuis le cabinet médical, les données sont transférées à un centre de calcul afin de procéder à leur analyse. Cette étape d'analyse des flux audio et vidéo se termine par une indexation de la vidéo selon les activités ciblées par le médecin. Les données indexées sont alors renvoyées au cabinet médical.

Depuis son cabinet, le docteur peut alors visualiser un panel d'activités exécutées par le patient à domicile. La segmentation automatique permet une navigation facile et rapide dans la vidéo. La visualisation de ces activités s'ajoute alors aux autres modalités du diagnostic, comme les entretiens, et permet de préciser le diagnostic par un regard plus précis sur la situation du patient dans son environnement habituel.

FIGURE 1.8 – Schéma général du projet IMMED.

1.3.4 Historique et objectifs du projet

Le projet IMMED a pris la suite d'un projet exploratoire « Monitoring Vidéo Embarqué » financé dans le cadre d'un appel PEPS du département S2TI du CNRS. Ce projet, premier en France à investir le champ de la caméra portée pour une application médicale, a permis de préciser des contraintes sur le dispositif d'acquisition, telles que la position et le type de caméra.

Le projet IMMED ANR-09-BLAN-0165-02 a été financé pendant trois ans par l'Agence Nationale de la Recherche (ANR) dans son programme Blanc. Il a impliqué trois partenaires académiques dans les STIC ainsi qu'un centre de recherche en santé publique :

- le LABRI (UMR 5800 CNRS, Université Bordeaux 1),
- l'IMS (UMR 5218 CNRS, Université Bordeaux 1),
- l'IRIT (UMR 5505 CNRS, Université Paul Sabatier),
- et le Centre INSERM U897 Epidémiologie et Biostatistique.

Le projet a commencé en septembre 2009 avec les objectifs suivants :

- un dispositif portable d'acquisition audio/vidéo qui respecte les contraintes ergonomiques de l'application médicale,
- des méthodes d'analyse automatique de flux audio et vidéo afin de segmenter automatiquement la vidéo en activités réalisées par le patient,
- une interface de visualisation de la vidéo structurée en activités,

– une validation des technologies par leur intégration dans une étude clinique et la définition d'un guide de diagnostic adapté à ce nouveau paradigme.

1.4 Réalisation du corpus IMMED

Dans le cadre de cette thèse, en partie financée par le projet IMMED, nous avons travaillé sur le développement de méthodes d'analyse automatique de flux audio pour la reconnaissance d'activités en milieu domestique. Il convient dans un premier temps de préciser le système d'acquisition des données et les caractéristiques du corpus recueilli.

1.4.1 Système de capture audio-vidéo

FIGURE 1.9 – Dispositif d'acquisition audio-vidéo.

Un système de capture audio-vidéo, permettant de filmer les activités instrumentales du patient, a été développé dans le cadre du projet. Ce dispositif, illustré à la figure 1.9, a été réalisé à partir d'une caméra portative HD-GoPro Fisheye fixée sur un gilet porté par le patient. La figure 1.10 nous montre des activités de la vie quotidienne filmée par le dispositif d'acquisition. Ces activités ont été effectuées par un volontaire sain et font partie du corpus IMMED.

FIGURE 1.10 – Images extraites du corpus IMMED.

Ergonomie

Les tests réalisés via des questionnaires sur des volontaires âgés sains ainsi que sur des patients souffrant de troubles légers n'ont pas mis en évidence de gêne des personnes vis-à-vis du dispositif.

Il semblerait que le dispositif puisse être rapidement oublié par le porteur de la caméra qui a tendance à se focaliser sur son environnement extérieur.

Sortie vidéo

La sortie vidéo est de bonne qualité. La caméra utilisée produit 25 images par seconde de dimension 1280 x 960 pixels. Le placement du dispositif sur l'épaule du patient et l'objectif Fisheye permettent à la fois d'observer les mains du patient pendant les activités mais aussi une grande partie du contexte extérieur.

Sortie audio

Au niveau audio, la caméra GoPro possède un micro unique. Le signal audio enregistré dans la vidéo est un signal mono échantillonné à 48 kHz. Le dispositif GoPro étant à la base dédié aux sports extrêmes, il utilise des algorithmes internes pour diminuer le bruit du vent. Nos contacts avec la société qui commercialise ces caméras ne nous ont pas permis de savoir précisément quels algorithmes sont utilisés par la caméra.

L'analyse des prises de son révèle néanmoins du repliement spectral, comme nous le voyons sur le la figure 1.11 qui représente un spectrogramme calculé sur un extrait du corpus IMMED. Nous pouvons voir une brutale coupure dans les fréquences autour de 15,3 Hz, alors que la fréquence d'échantillonnage est de 48 kHz (ce qui implique une bande passante potentielle allant de 0 à 24 kHz). Une analyse plus fine nous permet d'observer le repliement spectral : la portion fréquentielle présente entre 6 kHz et 12 kHz est la symétrie, atténuée, de la portion de fréquence compris entre 0 et 6 kHz. De même la portion présente au-dessus de 12 kHz est également une symétrie des fréquences inférieures.

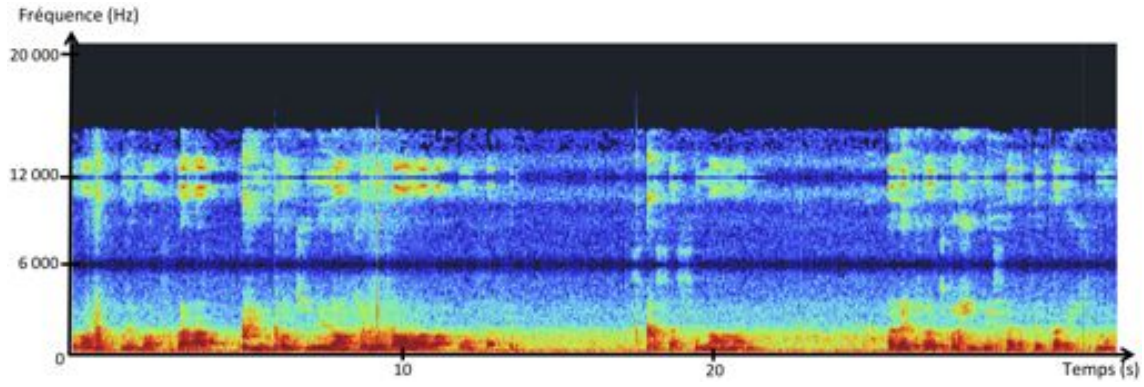


FIGURE 1.11 – Repliement spectral sur un spectrogramme d' un extrait audio de 30 secondes.

Le constat de ce repliement spectral nous a poussés à n'utiliser que la partie du signal comprise entre 0 et 6 kHz. La méthode utilisée pour filtrer le signal sera détaillée dans le chapitre 2.

1.4.2 Description du corpus

Vidéos enregistrées

Au cours des trois années du projet IMMED, de nombreuses vidéos ont été réalisées. Ces vidéos sont filmées grâce à la caméra portée par des patients ou des volontaires sains, qui réalisent

des activités de la vie quotidienne dans leur domicile. Le tableau 1.1 détaille le nombre et la durée des vidéos effectuées sur la durée du projet.

	Nombre d'individus	Nombre de vidéos	Durée
Volontaires sains	12	15	7h16
Patients	42	46	17h04
Total	54	61	24h20

Table 1.1 – Vidéos tournées lors du projet IMMED.

Particularité de la caméra portée

Le projet IMMED se différencie fondamentalement de ses prédécesseurs par l'utilisation des flux audio et vidéo issus de la caméra portée. Comme nous l'avons vu dans la partie 1.2, si les études de l'état de l'art utilisent régulièrement des capteurs audio et vidéo, elles les emploient plutôt de manière fixe. La captation par une caméra portée implique des conséquences importantes sur le corpus, au niveau vidéo et audio. Nous allons voir comment ces caractéristiques compliquent la tâche de reconnaissance automatique d'activités à partir du son.

1.5 Hétérogénéité des données

1.5.1 Reconnaissance d'évènements sonores

Nous avons commencé notre travail de reconnaissance automatique d'évènements sonores par une écoute attentive des enregistrements du corpus IMMED. À partir des activités d'intérêt, nous avons identifié des types d'actions présentant des caractéristiques sonores particulières. Réciproquement, l'existence de certains sons nous permet de deviner l'activité que le patient est en train de réaliser, par exemple lorsqu'ils proviennent de l'utilisation des objets suivants :

- porte,
- téléphone,
- télévision ou radio,
- robinet,
- livre ou magazine,
- machines à café,
- aspirateur,
- four.

Certains de ces sons peuvent être facilement reconnus à l'écoute. Pour d'autres, il semble nécessaire de regarder la partie vidéo des enregistrements. Dans tous les cas, leur reconnaissance automatique dans des contextes variés semble difficile.

Tout d'abord, certaines activités comme la lecture ou l'écoute prolongée de la télévision ou de la radio ne sont pas très présentes dans le corpus IMMED, l'assistant médical ayant tendance à solliciter le patient afin qu'il effectue des actions. D'autre part, les particularités du projet rendent difficile la mise en place de systèmes de reconnaissance automatique.

1.5.2 Conséquences des particularités du corpus

Pluralité des lieux de tournage

Les vidéos du corpus IMMED sont tournées au domicile des patients. Ainsi tous les films proviennent d'habitations différentes, mises à part certaines vidéos qui ont été découpées en deux lors de l'acquisition pour des raisons techniques. Les objets et les bruits impliqués dans les différentes activités peuvent différer profondément d'un lieu à l'autre. Par exemple, les sons liés à l'utilisation d'une machine à café varient selon les dispositifs et les manières de préparer le café. Cette variation importante rend difficile la reconnaissance d'activités à partir d'évènements acoustiques.

De plus les enregistrements sont aussi « dégradés » par l'environnement sonore autour de l'habitation. La présence d'une route, le passage d'avions, ou même le chant des oiseaux rajoutent des sources sonores qui, en fonction des maisons, compliquent la reconnaissance automatique, en masquant les sons à reconnaître ou en créant de fausses alarmes.

Enfin, les différents lieux d'enregistrement engendrent sur les sources sonores des réverbérations différentes. Par exemple, un objet utilisé dans une salle de bain carrelée ne produira pas le même enregistrement sonore dans un salon équipé de moquette et d'étagères. Chaque lieu crée ainsi un espace acoustique différent dont l'impact sur le signal enregistré peut avoir des conséquences importantes sur la reconnaissance de la source sonore.

Caméra mobile

Une autre difficulté réside dans le fait que la caméra est portée par les patients. Cette caméra mobile est donc soumise à ses mouvements. S'il est évident que la reconnaissance automatique à partir de l'image du signal est délicate du fait de prises de vues non maîtrisées et des mouvements de la caméra, les mouvements pendant l'enregistrement altèrent également la prise de son. Celle-ci est également dégradée par des chocs sur la caméra.

De plus, la position du microphone n'est pas choisie, comme cela peut être le cas dans d'autres études impliquant des microphones fixes. La position du micro mobile dépend de celle du patient, qui n'est pas toujours idéalement placé par rapport à la source à enregistrer.

Présence de parole

Dans les vidéos du projet, un assistant médical est présent pendant toute la durée de l'enregistrement. Cet assistant médical engage généralement une conversation avec le patient et lui indique les activités à effectuer. La conversation est souvent ravivée à chaque nouvelle activité proposée. La parole est donc extrêmement présente dans le corpus IMMED : d'une part la voix de l'assistant médical, d'autre part celle du patient. Du fait de la position de la caméra embarquée, près de la bouche du patient, la voix de ce dernier est très forte par rapport aux autres sons. Le micro peut même saturer quand le patient parle fort.

La présence importante de voix est donc très problématique pour la reconnaissance de bruits d'intérêts qui pourraient nous permettre de détecter les activités du patient, dans la mesure où elle peut sensiblement recouvrir ces bruits d'intérêts.

En conclusion, certaines caractéristiques du corpus compliquent fortement la tâche de reconnaissance. Il est assez clair que la reconnaissance automatique d'activités à partir de la vidéo constitue déjà un défi scientifique d'envergure dans un environnement maîtrisé avec des capteurs

fixes. Dans le contexte du projet, la pluralité des lieux, la mobilité du microphone, et la forte présence de parole impliquent une très grande hétérogénéité des évènements sonores à reconnaître. Cette diversité rend très difficile la conception d'un système de reconnaissance automatique robuste basé sur une modélisation des évènements.

1.5.3 La place des sons d'eau et des sons d'aspirateur

Face à cette difficulté, nous avons relevé deux évènements sonores régulièrement présents dans le corpus dont la détection pouvait être concrètement envisagée : les sons liés à l'utilisation d'un robinet (dans le cadre de la cuisine, de la salle de bain, ou autre) et ceux liés à l'utilisation d'un aspirateur.

Les sons d'aspirateurs ont l'avantage d'être assez bruyants pour ne pas être recouverts par d'autres sons, et d'être présents sur des plages temporelles assez longues.

Les sons liés à l'utilisation d'un évier sont présents sur un nombre important d'activités d'intérêt pour les médecins : dans les activités d'hygiène (se laver les mains, se laver les dents), les activités d'entretien (faire la vaisselle, faire le ménage, arroser les plantes) et les activités de cuisine (préparer un repas, une boisson).

De plus, les sons d'eau ont l'avantage d'être relativement universaux par rapport à d'autres bruits. En effet, si les objets impliqués dans les interactions acoustiques, par exemple les éviers, peuvent différer, les sons résultant de ces interactions sont plutôt homogènes.

La détection des sons d'eau et des sons d'aspirateur semble ainsi envisageable de manière robuste dans le contexte difficile du projet IMMED. Les informations sur l'utilisation de l'eau et de l'aspirateur peuvent nous permettre d'inférer sur les différentes activités que le patient est en train d'effectuer. Nous allons voir comment la construction d'un système robuste d'identification des sons d'eau pouvant fonctionner dans n'importe quel lieu s'est révélé être un défi scientifique suffisamment riche pour y consacrer la majeure partie de cette thèse.

1.6 Conclusion

Depuis plusieurs années, les technologies de l'information et de la communication sont utilisées régulièrement dans des applications médicales, comme par exemple au sein des maisons intelligentes. Dans ce chapitre, nous avons vu que le diagnostic des démences telle que la maladie d'Alzheimer peut être amélioré par le suivi des activités quotidiennes du patient effectuées à domicile. Le projet IMMED, premier en France à utiliser la caméra portée, propose d'enregistrer les activités des patients dans leur propre domicile. Une étape de segmentation automatique en activités est nécessaire pour faciliter la consultation des vidéos par les médecins.

Nous avons principalement travaillé au sein de ce projet sur la détection d'évènements sonores pour la reconnaissance d'activités. Les difficultés liées aux caractéristiques du projet nous ont alors amenés à nous focaliser sur la reconnaissance de sons spécifiques, dont les sons d'évier et les sons d'aspirateur. Les sons d'évier ont l'avantage d'être présents dans plusieurs activités d'intérêt pour les médecins, et sont assez homogènes selon les différentes habitations.

Des algorithmes de reconnaissance de sons liés à l'eau ont ainsi été développés. Le chapitre 2 dresse un inventaire général des applications et des méthodes utilisées dans le cadre de la reconnaissance automatique de sons environnementaux.

La création d'un système de reconnaissance automatique de son d'eau pour des applications médicales s'est révélé être un problème déjà abordé dans la littérature scientifique. Nous le verrons dans le chapitre 3, qui décrit précisément les systèmes de reconnaissance de « flux » d'eau.

Nous verrons de plus dans les chapitres 4 et 5 que la reconnaissance de sons d'eau est une problématique assez riche pour mériter d'être abordée selon les domaines de l'acoustique et de la perception.

Cette étude nous semble d'autant plus pertinente que les sons liés à l'eau sont intemporels et sont régulièrement entendus dans la vie courante, au sein et en dehors du domicile. S'attacher à construire un système pour reconnaitre les sons d'eau automatiquement, ou à comprendre comment les humains les identifient constitue donc une problématique passionnante, qui, si elle peut être directement appliquée au sein du projet IMMED, dépasse largement son cadre.

Chapitre 2

Le monde sonore : observations et techniques

Résumé du chapitre: L'audition humaine présente une faculté à identifier les sources dans un mélange sonore de la vie quotidienne. Le travail de recherche sur la reconnaissance automatique de ces sources est pourtant majoritairement orienté sur l'analyse de sources isolées. Depuis plusieurs années l'analyse computationnelle de scènes sonores pose les bases théoriques de la reconnaissance de sources dans un mélange. Nous présentons dans ce chapitre un état de l'art des techniques utilisées pour détecter un ensemble de sons, ou un son unique, dans un mélange de sources sonores.

2.1 Observations et définitions

2.1.1 Des paysage sonores

Nous sommes confrontés dans notre environnement quotidien à une multitude de sons. Ces sons peuvent provenir de différentes sources, par exemple le bruit du vent, le chant des oiseaux, les discussions de voisins, les bruits de pas, de machine, de véhicules ou encore d'avions...

La description de notre environnement sonore a été formalisée selon un point de vue historique et social dans une étude effectuée par Murray Shaffer, qui a proposé la notion de « paysages sonores » traduit de l'anglais *soundscapes* [Sch77]. Dans ces paysages sonores, Shaffer distingue plusieurs catégories de sons :

- les sonorités maîtresses ou toniques (*keynote sounds*). Ces sons désignent le bruit de fond d'un lieu, qui passe souvent inaperçu,
- les sons à valeur signalétique ou signaux sonores. Ils apportent une information aux auditeurs et renvoient souvent à une représentation ou à une cause,
- les marqueurs sonores. Ils sont emblématiques d'un lieu et servent à en définir certaines qualités qui le rendent remarquables.

Par ailleurs, la notion de paysage sonore telle que formulée par Shaffer est indubitablement reliée à l'individu qui l'écoute ou le perçoit.

2.1.2 Mélange de sources sonores

La perception auditive est ainsi un moyen de construire une représentation utile du monde qui nous entoure. Cette activité d'écoute est effectuée continuellement de manière plus ou moins consciente. La construction d'une représentation de l'environnement par l'écoute n'en demeure

pas moins une tâche difficile. Le principal problème, soulignée par Helmotz au milieu du 19e siècle, est que les sources sonores s'additionnent pour former un mélange sonore.

In the interior of a ball room [...] we have a number of musical instruments in action, speaking men and women, rustling garments, gliding feet, clinking glasses, and so on [...] a tumbled entanglement [that is] complicated beyond perception. And yet, [...] the ear is able to distinguish all the separate constituent parts of this confused whole [Hel63].

Lors d'un tel mélange sonore, la localisation et la compréhension d'une source précise peut être soumise, au moins partiellement, à un phénomène de masquage, produit par des sources dont les zones temps/fréquences sont trop proches. Un exemple célèbre de ce mélange de sources dans le cadre de la parole, est le *cocktail party problem*, décrit par Cherry dans les années 50 [Che57]. L'effet *cocktail party* reflète ainsi la capacité du système auditif à sélectionner une source sonore dans un environnement bruyant, tout en restant réactif aux autres sources.

2.1.3 Analyse de scènes sonores : approche perceptive

Différentes recherches sur la compréhension de ces mélanges sonores ont été synthétisées par Albert Bregman dans son livre *Auditory scene Analysis* paru en 1994 [Bre94]. Dans cet ouvrage, Bregman souligne de nombreuses similarités entre l'audition et la vision. Il formalise la perception de scènes sonores et identifie deux phases :

- la première est une phase de segmentation dans laquelle le flux audio est découpé en une collection de zones temps/fréquences locales, appelées segments,
- la seconde est une phase de regroupement dans laquelle les segments qui semblent provenir de la même source sont associés.

Cette seconde phase de regroupement peut s'effectuer sans connaissance *a priori*, lors d'une stratégie de type *bottom-up*. Dans cette stratégie, les zones suffisamment proches selon des critères temporels ou fréquentiels sont fusionnées de telle sorte que la perception globale précède les détails. Cette forme globale permet alors d'établir une représentation sémantique de la source. Selon Bregman, cette stratégie de perception, dite primitive, possède des analogies avec les propriétés de perception visuelle formulées par la théorie de la Gestalt [Kof35].

Par exemple, dans la perception visuelle, les objets se déplaçant à la même vitesse sont regroupés. La figure 2.1 montre un autre exemple classique de la théorie de la Gestalt où la forme globale précède les détails : les formes du vase ou des visages apparaissent selon une perception multi-stable, et prennent le dessus sur la perception des détails de la figure.

FIGURE 2.1 – Le vase de Rubin.

Dans une autre approche de regroupement exposé par Bregman et dite « basée sur des schémas », les connaissances *a priori* de l'auditeur vont l'aider à regrouper les segments appartenant à la même source. Ainsi nos connaissances sur la parole nous aident à regrouper les phonèmes

d'un discours. De même la connaissance de notre prénom nous permet de le discerner facilement dans le bruit.

2.1.4 Vers l'analyse automatique

Historique

La reconnaissance de contenu sonore par des machines est un domaine scientifique qui sévit depuis les années 1950. À l'origine, les travaux portaient principalement sur la parole, la tâche principale étant la transcription automatique de la parole. Un système développé dans les laboratoires BELL permettait ainsi de reconnaître 10 chiffres isolés en 1952 [DBB52]. Les systèmes de transcription automatique ont connu depuis des améliorations continuelles.

Les travaux des laboratoires BELL portaient également à cette époque sur la musique, avec notamment les travaux de Max Mathews et les premières pièces de musiques numériques. Au niveau de l'analyse, la transcription automatique d'une pièce musicale est une tâche dont la difficulté n'a rien à envier au domaine de la parole. Il est souvent nécessaire d'analyser plusieurs sources simultanément, la musique impliquant régulièrement plusieurs instruments ou plusieurs notes au même instant. Dans le cas de la parole comme dans celui de la musique, la question de la transcription de plusieurs sources simultanées reste encore aujourd'hui un problème ouvert.

Depuis les années 90, avec l'ère du multimédia et l'avènement d'internet, de nouvelles problématiques sont apparues, notamment celle de la recherche de documents dans des « masses de données ». Cette problématique semble aujourd'hui la manière la plus classique d'introduire une contribution scientifique dans le domaine. Les collections de documents sonores ont ainsi contribué à développer de nouvelles tâches, qu'il a fallu automatiser compte tenu du volume de données. Nous pouvons citer par exemple la détection de reprises musicales parmi un corpus de chansons.

Type de corpus utilisé

Récemment, de nombreuses tâches de reconnaissance automatique audio ont été évaluées dans un grand projet à vocation européenne, le projet Quaero [Qua13], dont l'IRIT est partenaire. De nombreux algorithmes ont été évalués sur les différentes tâches. Dans ce projet, nous pouvons remarquer que la plupart des évaluations ont été effectuées sur des corpus que nous appellerons « maîtrisés ». Ce phénomène n'est pas une particularité de ce projet et se retrouve dans un grand nombre de publications scientifiques actuelles sur la reconnaissance sonore.

En effet, la plupart des efforts dans ce domaine portent sur l'analyse de sources isolées et engistrées dans des conditions maitrisées, ou sur l'analyse de mélanges sonores produits à partir de ces mêmes sources. Par exemple, les tâches impliquant le locuteur sont régulièrement effectuées sur des corpus radiophoniques ou télévisuels, dont les enregistrements ont été effectués en studio, sans bruits « parasites ». La musique provient elle aussi de productions effectuées en studio, de façon parfois très formatée comme le corpus RWC de musique pop japonaise, qui reste malgré tout très exploité dans le domaine *Music Information Retrieval* [GHNO02].

Il semble clair que la problématique du mélange de source sonore, telle que décrite plus haut par Helmotz (partie 2.1.2), n'est pas l'axe principal des recherches effectuées dans le domaine de la compréhension automatique du monde sonore.

Séparation de sources

En parallèle de ces travaux d'interprétation, d'autres recherches ont pour objectif la séparation des sources constituant le mélange sonore. Plusieurs méthodes semblent ainsi émerger, comme celle de la formation de faisceaux (*beamforming*) [BW01], ou l'Analyse en Composantes Indépendantes (ACI) [BS95].

Une alternative notable réside dans le débruitage, dans laquelle le signal est vu comme la combinaison d'un signal utile et d'un bruit. Des recherches portent ainsi sur le débruitage de la parole, ou *speech enhancement* [RP13]. Les taches de transcription automatique dans un environnement bruité trouvent par exemple des applications dans la robotique.

D'autres techniques s'appuient sur la déréverbération qui permet de supprimer l'effet de salle d'un enregistrement et de faciliter la compréhension [YSD+12].

La séparation de sources et le débruitage sont toutefois limités par rapport à la perception de l'effet « cocktail party » par un humain. La séparation de sources requiert en général un nombre important de microphones, alors que la perception auditive s'effectue seulement grâce à deux récepteurs, les deux oreilles. Cette approche semble donc assez éloignée de l'analyse de scènes sonores tel que Bregman l'a énoncée. Dans le cas du débruitage, des hypothèses très restrictives sont souvent effectuées sur les sources, comme l'hypothèse de stationnarité.

2.2 Analyse computationnelle de scènes sonores

2.2.1 Principe

Dans le cas où de nombreuses sources se superposent de façon difficilement prédictible (par exemple dans la scène sonore de bal décrite par Helmotz dans la partie 2.1.2), il semblerait que les travaux de modélisation de contenus audio soient restés minoritaires. Pourtant, grâce à l'amélioration des techniques utilisées sur les sources isolées, ce domaine semble en expansion depuis les années 2000.

Ce domaine de recherche a été baptisé *Computational Auditory Scene Analysis*, ou CASA, en référence à l'analyse de scènes sonores décrite par Bregman. Un ouvrage de DeLiang Wang and Guy J. Brown paru en 2006 dresse un état des lieux des objectifs et des techniques [WB+06]. Le but de l'analyse computationnelle de scènes sonores est d'arriver à une compréhension d'une scène sonore, dans des performances comparables à celles d'un humain. Wang propose ainsi de réduire les corpus utilisés dans ce domaine à des enregistrement contenant au plus deux canaux.

La difficulté exposée dans le domaine CASA découle du recouvrement éventuel entre plusieurs sources. Selon Temko, ce recouvrement de sources sonores serait ainsi à l'origine de 70% des erreurs dans la campagne d'évaluation CLEAR [TN09] (pour plus de détails voir la partie 2.4.4).

2.2.2 Sons environnementaux

A la différence des travaux de reconnaissance exposés précédemment (partie 2.1.4), les travaux effectués dans le domaine CASA s'appuient principalement sur des enregistrements effectués dans la vie de tous les jours dans des conditions non maîtrisées. Les scènes sonores issues de ces enregistrements font intervenir des sons particuliers, peu présents dans les corpus radiophoniques et télévisuels. Les sons qui nous intéresserent ici sont couramment appelés « bruits ». Ils peuvent être par exemple identifiés comme des :

- bruits de pas,
- coups de feu,

– chocs entre objets,
– passages de voitures ou d'avion,
– bruits de verre cassé,
– ou encore, comme le son de l'eau.

Ce type de sons n'est pas produit par de la parole, ni par de la musique. Nous les appellerons sons environnementaux en référence à une définition proposée par Gygi : *all naturally occurring sounds other than speech and music* [GKW07].

Le concept de son environnemental est abordé dans cette définition par le biais de la source sonore. Néanmoins, il faut signaler que la notion de son environnemental possède de multiples interprétations, un même son pouvant être classé de plusieurs manières. Ces considérations seront présentées plus en détail dans le chapitre 5.

2.2.3 Détection d'évènements sonores

Principe

La tâche *Audio Event Detection* est une sous-tâche de CASA. Elle propose de détecter un ou plusieurs évènements dans un mélange sonore. À chaque source est associé un label selon une liste de labels prédéfinis. Cette détection d'évènements sonores peut consister à identifier des sons environnementaux dans un mélange de sources.

L'utilisation d'un nombre important de labels prédéfinis rapproche cette tâche du problème général d'analyse de scènes sonores, au sens de la limite mathématique. Les travaux utilisant une liste importante de labels prédéfinis seront présentés dans la section 2.4.1.

D'autre études se focalisent sur la détection robuste d'un son unique dans un mélange de sources. Ainsi dans la problématique du projet IMMED décrite dans le chapitre 1, la détection du seul son produit par de l'eau peut nous permettre d'inférer sur l'activité que le patient réalise. De même, dans les applications liées à la sécurité, la détection robuste d'un coup de feu peut constituer une information très pertinente. Les applications de détection d'un son spécifique feront l'objet de la section 2.4.2.

Positionnement du domaine

Dans le domaine AED, les bruits à détecter sont assez spécifiques. Cette tâche est donc différente des approches proposant de segmenter le signal en parole/musique/bruit [LZL03, Pin04], la notion de bruit n'étant dans ce cas pas assez précise.

Il faut également souligner que des évènements sonores de label identique peuvent montrer des différences importantes au niveau acoustique dans le domaine AED. Cette tâche se différencie donc de la reconnaissance de jingle, qui consiste à détecter une séquence audio précisément définie acoustiquement.

De même cette tâche se distingue de la reconnaissance d'une séquence sonore précise dans un mélange sonore bruité ou effectué dans la vie réelle, tel que le propose l'application SHAZAM qui permet d'accéder aux métadonnées d'une musique diffusée dans des environnements bruités [Wan06]. Dans cette dernière application, les sources sonores à reconnaitre sont supposées très proches des originaux avec lesquels SHAZAM calcule des empreintes acoustiques. Ce dernier procédé n'est pas applicable dans de nombreux cas où la reconnaissance par des humains est triviale, par exemple pour les reprises musicales qui peuvent pourtant être facilement identifiées par la mélodie, l'harmonie, le rythme ou les paroles.

2.3 Techniques utilisées

Les techniques utilisées dans la détection d'évènements sonores sont globalement héritées de la reconnaissance audio et donc particulièrement du traitement de la parole. Par exemple dans le cas de la détection de fréquence fondamentale, un détecteur de fréquences fondamentales multiples se compose souvent d'un détecteur de fréquence fondamentale unique et robuste, et d'un séparateur de voix permettant de réitérer l'opération de détection plusieurs fois (pour plus de détails voir l'article de De Cheveigné sur le *multipich* dans [WB+06]).

Par ailleurs, ce sous-domaine de CASA est lié à la perception humaine et les techniques utilisées sont souvent inspirées du système auditif ou cognitif des êtres humains. Par exemple, certains descripteurs acoustiques, comme les coefficients cepstraux ou les gammatones, s'appuient sur un flitrage fréquentiel proche de celui effectué par l'oreille humaine.

Ce domaine s'inscrit dans le cadre de la reconnaissance de formes, dont l'un des enjeux est le choix des paramètres (ou descripteurs) extraits du signal. L'utilisation d'un descripteur permet de réduire le nombre infini de variations temporelles du signal dans un espace adapté à la compréhension de celui-ci. Ainsi, dans le domaine de la détection d'évènements acoustiques, l'enjeu est de trouver l'espace paramétrique qui permettra de discriminer chaque source à reconnaître. Un modèle de classification permet alors d'attribuer les segments de signal à une classe précise en fonction des descripteurs calculés.

Dans les parties suivantes nous allons présenter les méthodes de reconnaissance sonore les plus courantes. Nous présenterons d'une part les paramètres acoustiques, d'autre part les modèles de classification. La description mathématique de ces paramètres et méthodes est proposée respectivement dans les annexes B et C.

2.3.1 Les paramètres acoustiques

Paramètres classiques

Les descripteurs audio fournissent une information sur un caractère précis du signal. En général, ce dernier est découpé en trames d'analyse, de l'ordre de 40 ms, et une valeur de descripteur est calculée pour chaque trame. D'autres types de descripteurs existent également, par exemple le temps d'attaque d'un évènement sonore.

Dans une approche descripteur/modèle de classification, nous pouvons constater que la plupart des contributions scientifiques concernent principalement le modèle de classification et que les descripteurs calculés sur le signal restent globalement les mêmes. Plusieurs catégorisation de ces descripteurs classiques ont été proposées. Dans le projet CUIDADO [Pee04], ils sont organisés selon les thèmes suivants :

- paramètres de forme temporelle,
- paramètres temporels,
- paramètres d'énergie,
- paramètres de forme spectrale,
- paramètres harmonique,
- paramètres perceptifs.

La norme MPEG-7 a permis la standardisation de certains d'entre eux, par exemple les descripteurs bas niveaux (low levels descriptors, [KMS06]). Cette norme repertorie un ensemble de descripteurs classés selon la structure suivantes :

- descripteurs basiques (forme d'onde, puissance),

– descripteurs spectraux basiques (enveloppe spectrale, centroïde)
– descripteurs basiques du signal (fréquence fondamentale, harmonicité),
– descripteurs de timbre.

Dans cette thèse nous avons effectué un travail spécifique sur le choix des paramètres, décrit dans le chapitre 3. Nous proposons dans l'annexe B une description des paramètres utilisés, classés selon deux catégories : les paramètres temporels et les paramètres fréquentiels. La plupart de ces paramètres sont classiquement utilisés dans les études de reconnaissance audio. D'autres, comme les gammatones, semblent avoir prouvé leur intérêt plus récemment [VA12]. D'autres encore, comme le coefficient de variation [WW94] sont le fruit d'une recherche spécifique sur la description d'un aspect du signal.

Comparaison et combinaisons de paramètres

En général, les études de détection et de classification automatique utilisent un ensemble important de descripteurs. De plus, afin d'avoir une information sur leur évolution temporelle, les systèmes de classification audio prennent en compte leurs dérivées premières et secondes.

Une approche classique en apprentissage automatique consiste ainsi à produire une grande quantité de paramètres qui seront utilisés dans un classifieur, par exemple des mélanges de lois gaussiennes. Ce modèle de classification permet de donner plus de poids aux paramètres pertinents. Le point faible de cette approche est qu'il est difficile de savoir quel paramètre est pertinent et quelle information est réellement modélisée. De plus, cette approche est inadaptée lorsque les corpus d'apprentissage et de test sont trop différents.

Ce problème est d'autant plus délicat que le nombre de paramètres possibles est infini. Le laboratoire Sony a ainsi développé un algorithme appelé *Extractor discovery system* qui génère automatiquement un ensemble important de descripteurs à partir d'opérateurs basiques. Cet algorithme permet de trouver un paramètre pertinent à partir d'opérations élémentaires entre paramètres. Ainsi, dans [PR07], Pachet nous démontre sur un problème artificiel de classification, qu'un seul paramètre bien choisi (dans cette étude, le paramètre est choisi automatiquement parmi 40 000 possibilités) est plus apte à capturer l'information importante qu'une centaine de descripteurs « classiques ». Si cette approche semble limitée pour des raisons de dépendance au corpus et de surapprentissage, elle a l'avantage de nous monter que des paramètres simples construits « à la main » peuvent être bien plus robustes que d'autres généralement utilisés.

Une autre approche consiste à réduire la dimension de ces paramètres. L'Analyse en Composante Principale (ACP) est par exemple une méthode qui peut être utilisée avant la classification pour éliminer l'information la moins pertinente. L'Analyse en Composante Indépendante (ACI) permet également de réduire la dimension d'un ensemble de descripteurs en minimisant la dépendance statistique entre certains descripteurs.

D'autres études ont été menées afin d' évaluer la capacité d'un ensemble de paramètres à modéliser un problème particulier. Ainsi dans [MA09], les descripteurs de la norme MPEG-7 sont comparés au MFCC pour leur capacité à modéliser des sons environnementaux. Il semblerait que ces descripteurs obtiennent de meilleurs résultats que les MFCC dans cette étude, bien que le mélange des deux types de descripteurs donnent les meilleurs résultats. Par contre, dans le cas d'un problème général de reconnaissance sonore où peut intervenir de la parole, les MFCC semblent plus efficaces [KS04].

Les outils de calculs

Le calcul concret des descripteurs sur des fichiers audio est une opération régulièrement effectuée. Elle peut être coûteuse au niveau du temps de calcul. Il existe des outils de calcul pour compiler de nombreux descripteurs de manière efficace. Parmi eux, nous pouvons citer YAAFE (*Yet Another Audio Feature Extractor*), qui permet de compiler rapidement de nombreux descripteurs [MEF+10]. Cette boîte à outils, libre d'utilisation, a été développée par le laboratoire Telecom Paristech. D'autres boîtes à outils existent, comme par exemple la MIRtoolbox [LT07].

Par ailleurs, la visualisation de l'évolution temporelle du signal est aussi une approche très pratique pour évaluer la capacité d'un descripteur à modéliser un problème précis. Les plugins « vamp» permettent de programmer facilement un descripteur temporel ou fréquentiel, qui pourra être visualisé, avec d'autres descripteurs usuels, dans le logiciel Sonic Visalizer, développé par l'Université QueenMary [CLS10].

Conclusion

Il existe une très grande variété de descripteurs audio, mais seule les plus efficaces sont utilisés dans la majorité des études. Il semble pourtant que le choix de ces descripteurs puisse être déterminant pour la création d'un système automatique de reconnaissance sonore. L'utilisation de ces descripteurs dans un tel système est liée à un modèle de classification qui va permettre de classer des échantillons de signal suivant les valeurs de ces descripteurs.

2.3.2 Les méthodes de classification

Les paramètres acoustiques permettent de mesurer différents aspects du signal sonore. Dans une tâche de détection ou de classification, les labels sont attribués aux segments du signal en fonction de ces mesures. Cet aspect est couramment effectué par un modèle de classification. Ces modèles peuvent se différencier entre modèles statistiques (ou génératifs) comme les modèles de mélange de lois gaussiennes, et modèles discriminants, comme les machines à vecteur de support.

Nous présentons dans cette partie le principe général des méthodes classiquement utilisées en reconnaissance sonore. Une description mathématique de ces méthodes a également été produite dans l'annexe C. Nous décrivons ainsi les méthodes suivantes :

- les k-plus proches voisins,
- les modèles de mélange de lois gaussiennes,
- les machines à vecteur de support,
- les modèles de markov cachés,
- la factorisation en matrices non négatives.

Les k-plus proches voisins

La méthode des k-plus proches voisins, ou kNN pour *k-Nearest Neighbours* est sans doute la méthode de classification automatique la plus simple à appréhender. Pourtant, elle donne parfois des résultats très similaires (quoique légèrement inférieurs) aux méthodes plus élaborées que sont les Modèles de Mélanges de lois Gaussiennes (GMM) et les Machines à Vecteur de Support (SVM) (par exemple dans [TSGM10]). Cette approche de regroupement consiste à attribuer une classe à un échantillon selon la classe des k échantillons les plus proches.

Les modèles de mélange de lois gaussiennes

Le monde est il gaussien ? Les lois gaussiennes, ou normales, permettent de modéliser des phénomènes naturels issus de plusieurs évènements aléatoires. La loi normale est ainsi très adaptée pour certaines problématiques physiques comme la modélisation de la matière. Son utilisation s'est développée au 19e siècle jusqu'à la modélisation des comportements humains, et la conception de *L'homme moyen* d'Adolphe Quetelet, où elle a trouvé des limites. Elle semble également aujourd'hui peu appréciée des économistes, du fait de ses difficultés à modéliser des phénomènes comme les krach boursiers [11].

Les modèles de mélange de lois gaussiennes constitue néanmoins une des méthodes de l'état de l'art les plus utilisées en reconnaissance des formes. Elle est ainsi souvent employée à titre comparatif et sert de référence. Cette approche s'appuie sur l'hypothèse que les paramètres peuvent être modélisés par un ensemble de lois normales.

Les machines à vecteur de support

Les machines à vecteurs de support ou SVM pour *Support Vector Machine* sont devenues en quelques années une méthode de classification incontournable. Sur des problèmes de classification audio, elle semble en général donner des résultats légèrement supérieurs aux GMM. Elle permet de classifier des échantillons en deux classes.

En pratique, les SVM sont couramment utilisés pour des problèmes de classification multi-classes. Pour une classification à N classes, une approche courante consiste à utiliser N classifieurs SVM. Le i$^{\text{ème}}$ SVM est dans ce cas entraîné avec les échantillons de la i$^{\text{ème}}$ classe avec des labels positifs, et tous les autres échantillons avec des labels négatifs. Le résultat de ce système à N-SVM est la classe correspondant au SVM ayant la plus grande valeur [Vap98].

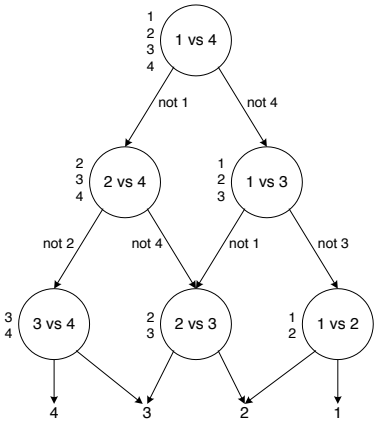


FIGURE 2.2 – Graphe orienté acyclique pour résoudre un problème de classification multi-classes.

Une autre approche consiste à étudier chaque combinaison de deux classes, et de construire C_N^k classifieurs SVM. Dans cette approche, il n'est toutefois pas toujours facile d'avoir une méthode fiable pour déterminer le résultat du classifieur global. Une des solutions peut être la construction d'un graphe orienté acyclique (Directed Acyclic Graph ou DAG), comme illustré à la figure D.4 [PCST99].

11. Causalité et finalité Par Gilles Cohen-Tannoudji

Les modèles de markov cachés

Les Modèles de Markov Cachés (MMC) constituent une technique extrêmement utilisée en segmentation et classification sonore. Cette technique a été particulièrement utilisée dans le traitement de la parole depuis la fin des années 60. Sa principale caractéristique, et son avantage par rapport aux techniques précédentes, est qu'elle modélise l'évolution temporelle des données. Ainsi, elle sera particulièrement utile pour reconnaitre des sons ayant une structure temporelle bien établie, comme par exemple les phonèmes dans la parole, ou pour modéliser l'enveloppe temporelle de certains sons.

La factorisation en matrices non négatives

La factorisation en matrices non-négatives, ou NMF pour *Non-negative Matrix Factorization* est aujourd'hui une méthode très utilisée pour la transcription polyphonique. La méthode mathématique de décomposition a été proposée à la fin des années 90 [LS99], mais c'est surtout depuis les travaux de Paris Smaragdis que cette méthode est très utilisée dans le domaine audio [SB03]. Le principe général de la NMF est de considérer le spectrogramme comme une matrice à coefficients positifs et d'établir une décomposition correspondant à la somme des sources sonores.

2.4 État de l'art

Nous présentons dans cette partie un état de l'art des applications du domaine *Audio Event Detection* (AED). Ces applications s'appuient majoritairement sur les techniques décrites précédemment et se divisent en deux types : les travaux proposant de reconnaître un nombre important d'évènements et ceux proposant la détection d'un évènement sonore unique.

2.4.1 Détection d'évènements multiples

Dans le contexte de la détection d'évènements multiples, un nombre important de sons peut être reconnu par le système. Nous proposons dans cette partie une classification selon le type de lieux : au sein du domicile, dans des environnements publics extérieurs ou intérieurs, et dans un cadre de bureau.

Détection à domicile

Les application liées à la détection à domicile interviennent généralement dans le domaine de la santé. Ce dernier domaine occupe une place privilégiée dans cette thèse puisqu' il a initié notre problématique de travail. Les applications liant la reconnaissance sonore au domaine de la santé ont ainsi été abordées dans le chapitre 1. Nous allons détailler dans cette partie les méthodes utilisées dans ces applications. À la différence de notre positionnement, ces méthodes cherchent en général à reconnaître un ensemble de sons d'intérêt au domicile des patients.

Vacher et ses co-auteurs proposent ainsi de reconnaître 7 classes de sons: vaisselle, porte, verrou, verre brisé, cris, téléphone et pas [VIB+03]. La maison est équipée de plusieurs microphones qui permettent d'avoir des prises de sons proches de l'évènement sonore. Dans ce système, la parole et les bruits mènent à des traitements séparés. La détection de cris provoque une alerte. Dans le cas de bruit, le système utilise une étape de détection suivie d'une étape de classification. Pour détecter les évènements saillants, il s'appuie d'une part sur l'autocorrélation du signal, un manque de corrélation permettant de détecter une nouveauté; d'autre part sur un système basé

sur les brusques changements d'énergie. Le système de classification utilise des mélanges de lois gaussiennes et un ensemble de paramètres, tels que les coefficients cepstraux, le *Zero Crossing Rate*, le centroïde spectral, le *Spectral Roll-Off*, les coefficients de prédictions linéaires, ainsi que leurs dérivées premières et secondes.

La suite de cette étude est détaillée dans un article très complet [ICV+06]. La détection de sons est ici améliorée grâce à une modélisation des sons d'impacts résultant d'une transformée en ondelettes. Toutefois, la plupart des expériences sont effectuées sur des corpus fabriqués artificiellement. Les résultats sur les données enregistrées dans la vie réelle ne sont pas publiés.

Dans une autre étude [PLST09], 7 évènements sonores sont détectés : tomber dans les escaliers, s'effondrer, crier, marcher, ouvrir et fermer une porte, sonnerie de téléphone, ainsi que la conversation. Le système proposé utilise un HMM hiérarchique appliqués sur les MFCC et ses dérivées. Les expériences sont effectuées sur des mélanges de sons créés artificiellement à partir de la base de données de sons environnementaux RWCP-SSD [NHA+00]. Elles montrent la supériorité du HMM hiérarchique sur les HMM simples. Le système obtient ainsi plus de 87% de performance globale. Toutefois, comme ces résultats s'appuient sur un corpus créé artificiellement à partir d'un ensemble de sons d'intérêt, il est difficile d'inférer sur les capacités d'un tel système à modéliser des enregistrements effectués dans des conditions réelles.

Détection dans des contextes variés

Dans [MHEV10], Mesaros propose de détecter 61 classes de sons à partir d'enregistrements de la vie réelle. Ces derniers enregistrements ont été effectués par son équipe de manière binaurale (avec un micro dans chaque oreille) dans différents environnements acoustiques (bus, match de basket, bureau, restaurant, plage, etc.). L'auteur propose de modéliser chacune des classes de sons par un HMM à 3 états utilisant des MFCC.

Dans une étude ultérieure, les résultats sur les mêmes enregistrements sont améliorés grâce à une prise en compte du contexte par une technique venant du traitement automatique des langues, l'analyse sémantique latente probabiliste. Grâce à cette méthode, il améliore son score de reconnaissance, passant de 30% à 35% de F-mesure.

Un autre type d'amélioration du même système est également proposé dans [HMVE11]. Dans cet article, Heittola et ses collaborateurs utilisent le résultat d'une séparation de sources effectuée par un algorithme de factorisation en matrice non négative. Le nombre de sources est fixé à 4, chiffre très inférieur au nombre de sources. Ainsi, l'algorithme ne peut trouver les sources séparées, beaucoup plus nombreuses, et les sources obtenues n'ont pas de sens d'un point de vue sémantique. Par contre, d'un point de vue computationnel, cette opération est suffisante pour simplifier le problème de détection et améliorer les résultats.

Détection dans des bureaux

D'autre méthodes ont été évaluées dans le projet CHIL [MMTN05] de la campagne d'évaluation CLEAR. Dans ce projet, l'une des tâches est la détection de sons spécifiques dans des enregistrements effectués lors de séminaires. Les contributions les plus importantes concernent toutefois la classification. Les évènements à reconnaître sont par exemple des bruits de porte, de pas, de téléphone, ou encore des rires ou des applaudissements. Temko propose une analyse des résultats des trois participants de la campagne dans [TMZ+06].

Le système UPC (Université Polytechnique de Catalogne) utilise un détecteur de silence, afin de les supprimer. La reconnaissance est basée sur une classification SVM étendue en multiclasse

grâce à un système de graphe orienté acyclique. Les descripteurs utilisés sont l'énergie par bandes de fréquence, le ZCR, le flux spectral, la fréquence fondamentale ainsi que leurs dérivées premières et secondes. Ce système utilise des fenêtres de 1 seconde avec 100ms de décalage.

Les deux autres systèmes, CMU (Carnegie Mellon University) et ICT (Istituto Trentino di Cultura), s'appuient sur des HMM et MFCC. Dans l'étape de détection, les deux premiers systèmes, UPC et CMU, effectuent segmentation puis classification, alors que le dernier (ICT), s'appuie sur un algorithme de Viterbi et obtient de meilleurs résultats. Une étude approfondie de ces systèmes et de leurs performances est disponible dans la thèse de doctorat d'Andriy Temko soutenue en 2007 et intitulée *Acoustic Event Detection and Classification*.

Dans une étude ultérieure [TN09], Temko propose de résoudre le problème de recouvrement de sources à la base de nombreuses erreurs dans le système UPC. Il construit ainsi une base d'apprentissage regroupant de nombreux cas de recouvrements. Cette astuce permet au système d'éviter 21% des erreurs, et d'arriver à presque 40% de réussite de détection.

Avec une autre approche, Harma propose de détecter les évènements « intéressants » afin de les classer, par l'intermédiaire d'un téléphone portable et d'un ordinateur de bureau [HMS05]. Les évènements intéressants se détachent du bruit de fond par leur énergie, ainsi que par la variance fréquentielle du pic d'énergie. Le bruit de fond est estimé par un lissage entre FFT consécutives. Pour la classification, les centres des classes sont déterminés manuellement. Des descripteurs usuels sont utilisés dans un classifieur kNN.

La campagne d'évaluation AASP Challenge [GBS+13] a récemment été proposée au sein de la conférence *Workshop on Applications of Signal Processing to Audio and Acoustics 2013*. Dans cette campagne, l'une des tâches était également la détection de sons spécifiques dans un contexte de bureau. À l'heure actuelle, les résultats et les techniques utilisées n'ont pas été diffusés.

Autres applications

Lifelogging Dans [AMSMH08], Shaikh se sert des sons environnementaux pour du *life logging* automatique. Le but de cette application est de récolter de manière passive et quasi permanente, des indices sonores sur la vie de l'usager, par exemple depuis un téléphone portable. Le but pour ce dernier, ou pour ces proches, est de garder une trace de toutes les activités effectuées. Selon Shaikh, les bruits du quotidien peuvent nous permettre d'inférer sur les activités réalisées. Il définit ainsi 114 types de sons, groupés dans 40 classes de sons, qui permettent d'inférer 17 types d'activités. Au niveau de la reconnaissance automatique, il utilise des HMM avec des MFCC. Les expériences sont menées sur des mélanges de sources synthétiques.

World Soundscape Project Le *World Soundscape Project* a été initié par Murray Shaeffer à la fin des années 60 dans un souci de sensibilisation à l'acoustique environnementale et de conservation du patrimoine sonore. Ce projet avait pour but de collecter d'immenses quantités d'enregistrements dans divers lieux. Le travail d'archivage étant considérable, on peut se douter que des outils d'annotation automatique faciliteraient grandement le travail d'archivage. La reconnaissance de sons environnementaux serait ainsi très utile pour un projet de ce type.

Un outil d'annotation semi-automatique a ainsi été présenté dans [KNA09]. Cet outil sélectionne automatique des zones temps/fréquences de forte énergie. La détection de ces zones est effectuée par une différence entre un cochléogramme et un cochléogramme lissé temporellement, où les détails sont masqués. Un utilisateur humain peut alors identifier manuellement ces zones

par un label. De plus le système propose automatiquement des labels pour les zones similaires à des zones déjà annotées.

Conclusion

Le parcours de ces différents projets de recherches nous permettent de dresser quelques conclusions pour le cas général de la tâche de détection d'évènements sonores :

- La tâche d'AED est très souvent séparée en deux phases : détection et classification. La tache de classification peut être testée seule sur des données déjà découpées, ce qui permet de tester les algorithmes sur des cas artificiels.
- Les scores sont souvent de l'ordre de 30 % de F-mesure, ce qui montre la difficulté de la tâche.
- Les expériences sont régulièrement menées sur des mélanges de synthèse, à cause de la difficulté de la tâche et pour des raisons pratiques. Il est alors assez difficile d'inférer sur les capacités du système à traiter des enregistrements issus de la vie réelle. Ce type de système semble ainsi difficilement applicable à notre projet IMMED.

2.4.2 Détection d'évènements spécifiques

Nous allons maintenant présenter quelques applications qui privilégient la robustesse de la détection au nombre de sons détectés. Ces applications proposent en général la détection d'un seul son. Nous proposons pour cette partie un classement des systèmes selon le type d'application : santé, écologie, sport et sécurité.

Santé

Dans les applications liées à la santé, certains systèmes utilisent le son d'eau pour la reconnaissance d'activités. Ces méthodes seront détaillées dans le chapitre 3.

Par ailleurs, une détection de chute à partir d'un microphone audio a été proposée dans [ZHPHJ09]. Pour détecter le bruit de chutes, 9 classes de sons sont modélisées, parmi lesquelles les sons correspondant aux actions *tomber*, *s'assoir*, *déplacer un objet* mais aussi les *bruits de pas*, ou encore la *parole*. Le système proposé utilise les coefficients de prédiction linéaire (LPC) ainsi que l'énergie du signal. Le modèle statistique utilise des GMM pour modéliser chacune des 9 classes, puis un SVM pour discriminer la chute des autres bruits. Les expériences sont effectuées sur un corpus acoustique de détection de chute dans un environnement quotidien, issu du projet Netcarity, et obtiennent 67% de F-mesure.

Ecologie

L'étude des sons des animaux est un domaine assez répandu en traitement du signal audio. Beaucoup de travaux portent ainsi sur la détection d'un animal spécifique. Les résultats peuvent être utilisés pour le comptage des individus d'une espèce précise afin de suivre son évolution.

Par exemple, Daniel Diep propose de détecter le passage de poissons migrateurs, les Aloses qui ont la particularité de donner des battements de queue sonore la nuit venue [DNM+13]. Le comptage au fil des ans permet de suivre l'évolution de l'espèce afin de prévenir son extinction. Dans un système implémenté sur un téléphone portable, des coefficients spectraux sont extraits du signal et classés par un GMM.

De nombreux travaux ont également été effectués sur les chants d'oiseau. Par exemple dans [BWK+10], l'auteur propose de repérer des formes temporelles dans des bandes de fréquences

typiques des espèces ciblées. D'autres applications existent, par exemple le suivi de population de baleines [LALM05].

Sport

Plusieurs études ont été effectuées sur ce sujet depuis le début des années 2000. Le but de cette application est de détecter un évènement sonore particulier permettant l'annotation ou la compréhension du sport.

Les impacts de balle ont ainsi été détectés, par exemple dans la cas du tennis de table [ZDC06]. Dans cette article, les auteurs proposent un système en deux phases, segmentation et classification. Il est intéressant de constater que ces deux phases utilisent des paramètres différents. Ainsi, la segmentation s'effectue par une détection des changements d'énergie brutaux dans le signal. La classification, par contre s'effectue de manière fréquentielle à l'aide d'une distance calculée sur les MFCC (par rapport à des échantillons de référence).

Dans une autre étude, dans le cas du tennis, on améliore la détection d'impact de balle en utilisant les cris des joueurs, qui parfois peuvent recouvrir les sons d'impact [HC12]. Les cris sont ainsi utilisés comme un évènement de contexte qui permet de déterminer la position des impacts de balle.

Sécurité

La sécurité est un thème qui bénéficie de nos jours d'une médiatisation très importante. C'est donc un enjeu des technologies de l'information et de la communication. Il apparait donc difficile de ne pas citer dans ce chapitre quelques exemples d'applications liées à la sécurité.

Ce thème remporte un tel succès que certaines entreprises se spécialisent dans ce domaine pour fournir des logiciels sur mesure qui équiperont potentiellement prison ou parking, en détectant par exemple les sons de vitres brisées [Ltd13]. En général, les évènements anormaux, comme les coups de feu ou les explosions, sont très puissants. Il peut être assez facile de les détecter. L'inventaire des approches possibles a ainsi été détaillée dans [Rex11].

Un système de surveillance a par exemple été proposé dans [CER05]. Des GMM sont utilisés pour la détection de coup de feu. Ce dispositif est implémenté dans un système multimodal qui utilise aussi la détection de cris.

Conclusion

Le domaine de la reconnaissance d'évènements spécifiques semble minoritaire par rapport à la reconnaissance générale d'évènements sonores. Ainsi, la spécialisation des applications est une difficulté pour la comparaison de système et de méthode, par exemple par la mise en place de campagne d'évaluation.

Les méthodes utilisées sont en général basées sur des algorithmes assez classiques. Toutefois il est intéressant de voir que l'adaptation de ces méthodes à des problèmes très précis amène une inventivité et une originalité qui permet d'améliorer sensiblement les résultats. À la différence des expériences du cas général, parfois très abstraites, la plupart de ces études sont effectuées sur des enregistrements de la vie réelle et affichent des scores qui rendent leur utilisation tout à fait envisageable dans un cadre complètement automatique.

2.4.3 Reconnaissance du contexte

Une des faiblesses de l'approche de détection d'évènements sonores pour la reconnaissance automatique d'activités ou de scènes complexes, réside généralement dans le manque de prise en compte du contexte. Un cri, par exemple, n'aura pas du tout le même sens pour un auditeur lors d'un jeu parmi d'autres cris que dans une rue silencieuse.

Il semble ainsi que la reconnaissance du contexte et celle des sons environnementaux qui le composent s'influence mutuellement. D'une part une bonne connaissance du contexte pourra nous permettre de reconnaître plus facilement les sons à venir. D'autre part la reconnaissance robuste de sons nous permet d'inférer sur le contexte. Il convient donc de citer quelques systèmes qui utilisent l'une ou l'autre de ces deux assertions.

La reconnaissance de scènes sonores en s'appuyant aidée par le contexte a fait l'objet d'une thèse de doctorat soutenue en 1996 par Daniel Ellis, intitulée *Prediction-driven computational auditory scene analysis* [Ell96]. Ainsi l'aide du contexte semble être une piste très intéressante pour certaines applications, comme celle liées à la sécurité. Toutefois, les techniques à mettre en œuvre sont très complexes et restent encore peu utilisées.

La seconde approche, assez classique dans le domaine CASA, est couramment énoncée comme reconnaissance du contexte. Une application classique est l'exemple du téléphone portable qui pourrait automatiquement passer du mode vibreur à la sonnerie en fonction du contexte (réunion, plein air, etc.). Dans la plupart des études de ce domaine, les évènements qui composent la scène sonore ne sont pas identifiés sémantiquement. Néanmoins, une approche basée sur des descripteurs et des classifieurs peut permettre d'avoir des résultats acceptables dans des enregistrements de la vie réelle [PTK$^+$02].

Il semble pourtant que ces approches négligent souvent l'aspect temporel des évènements sonores qui composent ce contexte. Dans [CNK09], Chu utilise un algorithme de Matching-Pursuit pour prendre en compte le domaine temporel. Des atomes de Gabor permettent de modéliser des évènements sonores dans le plan temps/fréquence. Les résultats sont comparés avec des tests perceptifs qui servent de référence.

2.4.4 Evaluation

Les campagnes d'évaluations CLEAR ou AASP Challenge ont été évoquées dans la section 2.4.1. Les métriques utilisées dans ces campagnes diffèrent sensiblement de celles classiquement utilisées en reconnaissance sonore.

En effet, dans un ensemble important d'applications de la reconnaissance sonore, l'aspect temporel est primordial. Par exemple, dans une segmentation en locuteur, il est important de trouver précisément les débuts et fins d'intervention de chaque locuteur. Le score de reconnaissance s'appuie donc sur la durée pendant laquelle l'évènement sonore, dans notre exemple le locuteur, a été correctement détecté.

Dans le domaine des sons environnementaux, les frontières temporelles peuvent avoir moins d'importance. Par exemple pour un détecteur de chute, il semble clair que la détection robuste de l'évènement est prioritaire par rapport à sa localisation temporelle précise. La campagne CLEAR propose ainsi de nouvelles métriques qui privilégient le fait qu'un évènement ait été reconnu par rapport aux frontières temporelles classiquement évaluées [SBB$^+$07].

Concrètement, dans la campagne CLEAR Audio Event Detection, un évènement détecté est considéré correct si son centre se situe entre les frontières d'au moins un évènement de la vérité terrain de même label, ou s'il existe au moins un évènement de même label dans la vérité terrain dont le centre est situé entre les frontières de l'évènement détecté.

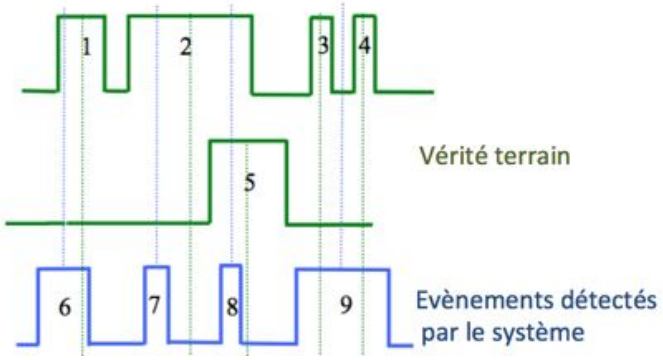


FIGURE 2.3 – Vérité terrain et évènements détectés.

Inversement, un évènement est considéré comme détecté pour des raisons similaires. Par exemple dans la figure 2.3, l'évènement 2 est considéré bien reconnu s'il possède le même label que 7 ou 8. De même, l'évènement 5 est considéré reconnu s'il possède le même label que 8.

Au delà de ces campagnes d'évaluation, les chercheurs sont régulièrement amenés à réaliser eux mêmes leur enregistrements. Ainsi, il est souvent impossible de comparer différentes méthodes à partir des publications car elles ne sont pas évaluées sur les mêmes corpus.

Il peut alors être intéressant de reprogrammer des méthodes pour connaître leur aptitude à modéliser un problème précis, bien que cela soit très coûteux en temps. Dautre part, les chercheurs pourraient dans l'idéal mettre à disposition les corpus enregistrés, comme dans le cas du projet *CIESS*[12].

2.5 Conclusion

Une des particularités fascinantes de l'écoute humaine est la faculté à sélectionner et identifier une source sonore dans un mélange de son. Cette faculté nous permet de fabriquer une représentation du monde qui nous entoure dans tout type de contexte. Les projets de recherche en reconnaissance sonore automatique se sont focalisés ces dernières décennies sur les sources isolées. Le problème de reconnaissance automatique de sons dans un environnement réel reste à ce jour un problème très délicat, même si les fondements théoriques ont été posés ces dernières années avec la thématique de recherche « Analyse computationnelle de scènes sonores ».

La détection de sons dans un mélange de sources sonores est une sous-tâche de ce domaine. Si différents travaux ont été effectués dans ce sens ces dernières années, il semble difficile d'appliquer une méthode unique pour reconnaître un nombre important de sons. Les travaux visant à détecter un son unique obtiennent des scores bien supérieurs qui permettent leur utilisation dans des conditions réelles. Dans cette thèse nous avons privilégié la robustesse et l'application pratique au nombre de sons à reconnaitre, en nous focalisant sur un seul type de son, le son provoqué par l'eau.

Nous avons également présenté dans ce chapitre les descripteurs et modèles statistiques les plus courants. Le système MFCC-GMM semble représenter un système classique de l'état de l'art que nous utiliserons dans la suite comme système de référence. Les gammatones sont une

12. http://www.petra.univ-tlse2.fr/ciess

alternative aux MFCC, plus proches de la perception humaine, que nous utiliserons également. Nous allons dans le chapitre suivant exposer notre méthode de reconnaissance de son de « flux » d'eau, en commençant par une sélection de descripteurs adaptés à cette tâche.

Chapitre 3

Reconnaissance de flux d'eau

Résumé : Divers travaux ont été effectués sur la reconnaissance automatique de sons d'eau, notamment dans le cadre des applications médicales. Ils sont néanmoins difficilement adaptables à notre problématique et à l'hétérogénéité des vidéos du projet IMMED. Nous présentons dans ce chapitre un nouveau descripteur audio, appelé couverture spectrale, qui permet de reconnaître les flux d'eau dans des environnements bruités. Des expériences effectuées sur le corpus IMMED sur plus de 7 heures de vidéo valident un système fondé sur de simples seuils. Notre système est ensuite amélioré par une étape de classification. Enfin, une expérience perceptive effectuée sur le même corpus révèle une certaine difficulté à définir, et donc à reconnaître, les sons d'eau.

3.1 Travaux antérieurs

La reconnaissance de flux d'eau pour des applications médicales a fait l'objet de plusieurs études antérieures à cette thèse. Ces travaux, principalement orientés vers le suivi à distance d'activités, témoignent de l'intérêt de reconnaître automatiquement l'utilisation de l'eau. Nous allons présenter ici de manière chronologique des études qui impliquent différents types de capteurs. Nous verrons également que le suivi de la consommation d'eau trouve des applications dans d'autres domaines, notamment en écologie.

3.1.1 Microphone dans la salle de bain

La première étude à notre connaissance dans le domaine de la reconnaissance d'activités à partir des sons d'eau date de 2005. Chen et ses collaborateurs proposent d'utiliser un microphone dans une salle de bain comportant douche et toilette (illustrée sur la figure 3.1), pour identifier les activités effectuées par une personne âgée [CKZ+05]. Par ce procédé, les auteurs proposent de suivre à distance les activités d'hygiène d'une personne âgée atteinte de trouble du comportement. Selon les auteurs, l'utilisation de ce suivi à distance évite des situations embarrassantes d'observation directe ou d'utilisation de la vidéo. Elle est aussi plus fiable que l'interrogation de patients souffrants de démence. Le système vise à reconnaître 7 types de sons, dont ceux liés aux activités : *se laver les mains, se laver les dents, se doucher, tirer la chasse d'eau.*

Les descripteurs audio du système sont calculés sur des fenêtres de 25 ms avec un recouvrement de 50%. Un vecteur constitué de 13 MFCC est extrait de chaque trame. La classification est effectuée par un modèle HMM. Chaque activité est modélisée par un HMM à densité continue de 6 états (sans sauts d'états possibles). Les états de cet HMM sont modélisés par un mélange de deux lois gaussiennes.

FIGURE 3.1 – Identification d'activités dans une salle de bain [CKZ+05].

Les expériences se sont déroulées pendant dix jours dans une salle de bain utilisée par quatre personnes. Quatre jours furent dédiés à l'apprentissage du modèle HMM. Les résultats sur les six jours restants révèlent des scores de reconnaissance supérieurs à 84% pour la plupart des activités.

3.1.2 Microphones dans les fondations de la maison

Dans un article de 2006, James Fogarty et ses co-auteurs proposent de suivre à distance les activités liées à l'eau dans une maison entière [FAH06]. Le principe est d'utiliser plusieurs capteurs sur les tuyaux d'arrivée et d'évacuation d'eau. Ce prototype a la vocation d'être non intrusif et peu coûteux.

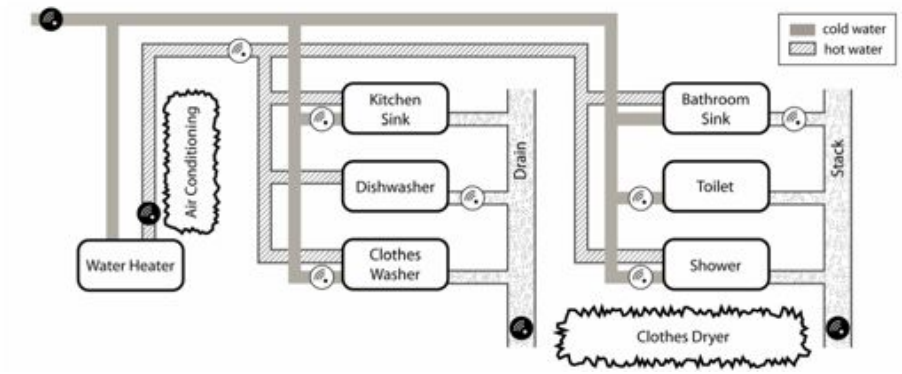


FIGURE 3.2 – Capteurs placés sur les tuyaux d'arrivés d'eau [FAH06].

Les capteurs utilisés sont des microphones associés à un système de transmission sans-fil. La figure 3.2 illustre les 4 capteurs de test (en noir) utilisés pour cette expérience, placées sur :

- l'arrivée d'eau froide,
- la sortie du chauffe-eau,
- l'évacuation de la cuisine et de la machine à laver,

– l'évacuation de la salle de bain et des toilettes.

Dans cette figure, nous pouvons aussi observer 7 autres capteurs de référence (en blanc) utilisés pour collecter les données de la vérité terrain.

Le système se base sur deux descripteurs sonores : le ZCR et l'énergie. Ces descripteurs sont calculés sur des extraits de 250 ms, et sont produits une fois toutes les deux secondes pour limiter la consommation d'énergie. À l'aide de ces descripteurs calculés sur les capteurs de référence, la vérité terrain est constituée : *utilisation du lavabo, de l'évier, des toilettes, de la machine à laver ou du lave-vaisselle.*

Au niveau automatique, un système basé sur un SVM permet de déterminer l'activation des capteurs de test. Un algorithme *ad hoc* permet de faire correspondre les séquences d'activation de ces capteurs de test aux activités effectuées. Cet algorithme permet par exemple de reconnaître les cycles d'utilisation de l'eau d'un lave-linge, qui produit une séquence d'activation précise sur le capteur correspondant.

Les données furent collectées pendant 6 semaines dans une maison occupée par deux adultes. La première semaine est utilisée en apprentissage et les 33 jours restants en test. Les résultats varient de 73% à 100% de reconnaissance pour les différentes activités. Toutefois les auteurs soulignent les difficultés rencontrées lors de l'utilisation simultanée de plusieurs points d'eau, par exemple la cuisine et la machine à laver. Ce problème avait déjà été mentionné par Chen dans son étude [CKZ+05].

3.1.3 Mesure du débit d'eau

Une autre étude propose un détecteur d'utilisation d'eau à partir d'un microphone placé sur le tuyau d'arrivée d'eau de l'évier de la cuisine [IBC+08]. L'objectif de cette étude n'est pas seulement de suivre l'utilisation du robinet de façon binaire, mais également de quantifier la quantité d'eau utilisée. Ce travail est toutefois plutôt axé sur la conception et la faisabilité de ce type de système et manque d'expériences pour être validé dans des conditions réelles d'utilisation.

Le système proposé permet de reconnaître le silence et 6 quantités d'eau différentes. Au niveau des descripteurs audio, les auteurs utilisent le logarithme des coefficients de Fourier filtrés par 31 filtres disposés selon une échelle Mel. Ce procédé correspond à une étape intermédiaire dans le calcul des MFCC. Une réduction de dimension est effectuée pour ne garder que 7 coefficients. Trois classifieurs sont testés sur un ensemble de test non bruité. Le modèle kNN donne des résultats supérieurs à ceux d'un arbre de décision ou d'un modèle SVM .

3.1.4 Détection de flot d'eau dans l'activité « se laver les mains »

Dans l'une des études les plus proches de notre problématique, Taati et ses co-auteurs proposent de détecter des flots d'eau à partir d'une caméra fixe située au dessus du lavabo [TSGM10]. Cette méthode est vouée à s'insérer au sein du système COACH [MBCH08], pour indiquer aux personnes âgées atteintes de la maladie d'Alzheimer les activités qu'elles ont réalisées.

Le système proposé utilise à la fois les données audio et vidéo issues de la caméra. Le signal audio est filtré par un filtre passe-haut qui supprime les fréquences en dessous de 4 kHz. Ce traitement permet de supprimer une partie des fréquences produites par la parole. Un rapport signal sur bruit est compilé par évaluation du bruit de fond. Un ensemble de descripteurs est ensuite calculé : ZCR, centroïde spectral, spectral Roll-off à 85%, flux spectral. Enfin, 19 MFCC sont également extraits du signal.

Au niveau vidéo, la détection de l'eau s'appuie sur les variations rapides des formes générées par l'eau. Une détection préalable du lavabo et un suivi des mains est utilisé pour identifier les

FIGURE 3.3 – Reconnaissance de l'activité « se laver les mains » [TSGM10].

zones où l'eau coule (voir figure 3.3), et en extraire des paramètres temporels.

Pour la reconnaissance, quatre techniques de classification avec apprentissage supervisé, dont les k-plus proches voisins et les SVM, furent testées. Les expériences furent menées sur 16 vidéos d'une durée moyenne de 2 minutes, avec un protocole *leave-on-out*. Les différentes méthodes obtiennent des résultats très proches : de l'ordre de 80% de détection correcte pour le système audio seul et 88% pour le système complet. Au niveau des résultats audio, les ajouts des descripteurs puis des MFCC au rapport signal sur bruit améliorent successivement les résultats. Toutefois, les résultats à partir des MFCC seuls ne sont pas présentés.

3.1.5 Détection de gaspillage d'eau

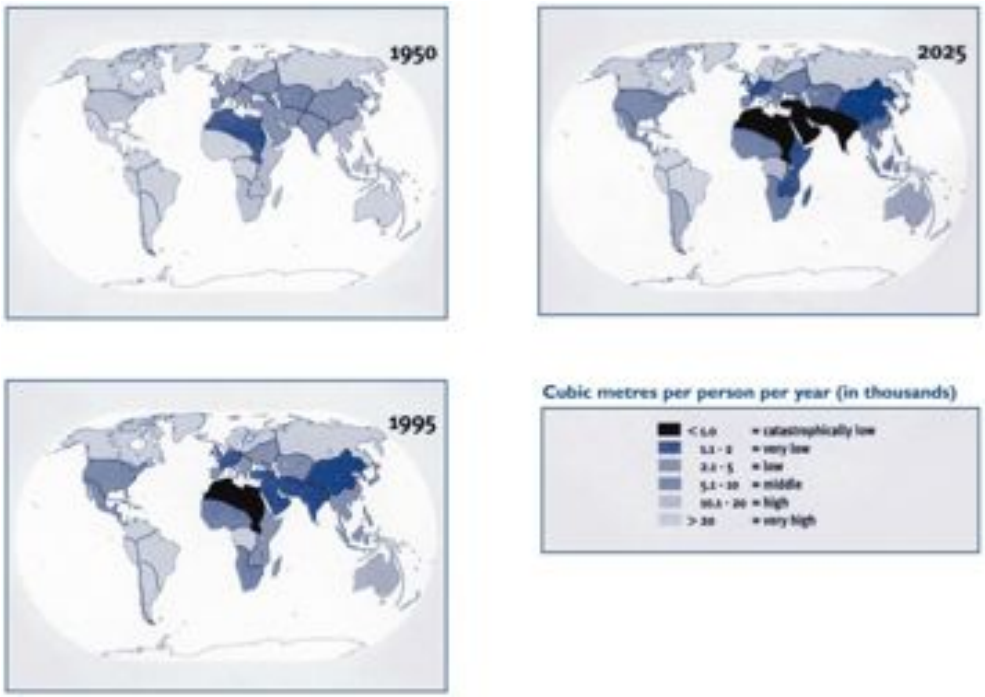


FIGURE 3.4 – Evolution de la consommation d'eau mondiale [Nat03].

Au delà de la reconnaissance d'activités, plusieurs études abordent le thème de la reconnaissance de l'utilisation d'eau au sein du domicile dans un objectif écologique lié à sa consommation.

Les rapides changements environnementaux, la diminution rapide des réserves dans certaines régions du monde et la difficulté à s'approvisionner en eau potable font de l'utilisation de l'eau un thème décisif dans la gestion des ressources naturelles. De plus, l'augmentation de la consommation mondiale, illustrée à la figure 3.4, laisse supposer un succès potentiel pour cette application dans les années à venir.

Détection de gaspillage

Dans [VSN+11], les auteurs proposent ainsi un détecteur de gaspillage d'eau à partir d'un microphone fixé sur le robinet. Le système vise à reconnaître les gaspillages d'eau entre les activités (inter-activités), par exemple lorsque le robinet n'est pas complétement fermé. Il permet également de détecter les gaspillages lors d'une activité (intra-activités) illustré à la figure 3.5 par l'utilisation du savon.

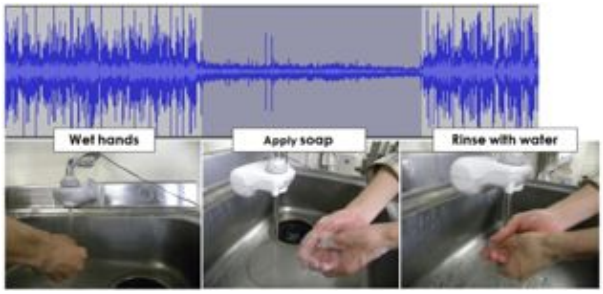


FIGURE 3.5 – Gaspillage d'eau lors de l'activité « se laver les mains » [VSN+11].

Le système proposé permet de détecter différentes classes d'activités : *se laver la figure, se laver les dents, se laver les mains, faire la vaisselle*, ainsi que les moments de gaspillage inter-activités et intra-activités.

Les enregistrements sont effectués à une fréquence d'échantillonage de 44,1 kHz. La détection de gaspillage et des activités s'effectue par une sélection très précise des coefficients issus d'une transformée de Fourier. Les coefficients sont calculés sur des fenêtres de 2^{13} échantillons (avec un fenêtrage de Hamming et un recouvrement de 50%). Une soustraction spectrale à partir des différentes activités permet d'identifier parmi les coefficients spectraux des descripteurs discriminants, par exemple l'amplitude minimale ou moyenne sur une bande de fréquence donnée. 40 descripteurs sont sélectionnés pour la classification inter-activités, et 35 pour la classification intra-activités. La classification s'effectue par un perceptron multicouche à trois couches (entrée, cachée, sortie).

Les expériences sont effectuées sur un corpus enregistré en laboratoire dans des conditions idéales et non bruitées. Les expériences affichent des résultats de l'ordre de 80% de détection correcte en moyenne. Il semble toutefois que les conditions d'enregistrement du corpus permettent de reconnaître précisément les différents types de sons mais ne sont pas réalistes pour une application dans la vie réelle.

Hydrosense

Le projet Hydrosense témoigne de la transversalité entre suivi de la consommation d'eau et suivi des d'activités de la vie quotidienne, en visant ces deux applications [FLC+09]. Cette étude démontre la possibilité de suivre l'utilisation et la consommation globale de l'eau dans une

maison, par l'intermédiaire d'un capteur unique. Ce capteur est fixé sur l'un des robinets (par exemple un robinet extérieur, voir figure 3.6), et permet d'enregistrer les variations de pression de l'eau dans la tuyauterie d'une maison.

FIGURE 3.6 – Détection des changements de pression dans le projet Hydrosense [FLC+09].

En général, la pression en eau au sein d'une maison est constante, sauf lors de l'ouverture ou de la fermeture d'un robinet qui entraîne une variation rapide. L'idée de ce projet est que le type d'appareil (évier, toilette, machine à laver), de valve, et la distance de tuyau les séparant du capteur confère à chaque appareil une signature unique qui peut être reconnue par les changements de pression dans la tuyauterie.

À partir de signaux de pression, quatre type de distance sont calculés : filtrage adaptatif, filtrage adaptatif dérivé, erreur quadratique moyenne, et filtrage adaptatif des coefficients cepstraux. Les coefficients ceptraux sont ici utilisés comme un modèle source/filtre où l'onde de pression originale est filtrée par sa propagation à travers les tuyaux de la maison. La classification est effectuée par des seuils.

Les expériences montrent des résultats très satisfaisants pour la détection des ouvertures et des fermeture des différentes valves. Un score de 97% de détection correcte est obtenu sur différents types de plomberie et de maison. Le système propose également une estimation de la quantité d'eau utilisée. Il semble robuste aux évènements simultanés, comme l'utilisation du robinet pendant le remplissage du réservoir de la chasse d'eau, car il permet de repérer chaque ouverture et fermeture de robinet.

Par contre, les expériences sont effectuées avec un utilisateur ouvrant chaque robinet de manière brusque et totale. Nous pouvons donc douter des possibilités d'un tel système à être utilisé dans des conditions réelles avec des personnes utilisant des débits d'eau variés selon leurs besoins. Ce système semble par contre très simple et pratique pour détecter l'utilisation des dispositifs à ouverture constante, tels que la chasse d'eau, la machine à laver ou le lave-vaisselle.

La mise en place pratique du projet Hydrosense a par ailleurs été améliorée par une contribution au niveau du dispositif de captation. Le système WATTR propose un système de captation de variations de pression auto-alimenté [CLC+10] et permet de s'affranchir de l'alimentation électrique du dispositif.

Show-me

Le système SHOW-ME vise à suivre l'évolution de la consommation d'eau pendant une douche [KG09]. Un affichage lumineux par LED permet de voir la consommation d'eau augmenter au fil du temps lors de l'ouverture du robinet. Les études à long terme effectuées dans des conditions réelles d'utilisation visent à démontrer une diminution de la consommation individuelle. La consommation d'eau est enregistrée à l'aide d'un compteur interne au tuyau de douche.

Ce système montre que d'autres applications à la reconnaissance de son d'eau sont possibles. L'usage d'un microphone permet en effet une utilisation plus souple, bien que probablement moins précise, que celle d'un compteur.

3.1.6 Conclusion

Les études décrites dans cette partie visent à reconnaître automatiquement, et éventuellement à quantifier, l'utilisation de l'eau au sein d'un domicile. Elles montrent une grande diversité d'approches dans le matériel et dans les algorithmes utilisés.

Certains travaux s'appuient ainsi sur le son provoqué par l'eau lors des activités. Toutefois, ils utilisent pour la plupart des méthodes génériques de reconnaissance sonore, appliquées cette fois au son d'eau, mais sans adaptation spécifique à ce type de son. Le travail mené par Taati fait exception puisque le filtrage passe-bas décrit dans cette étude permet de supprimer la voix du patient en gardant les hautes fréquences du spectre. Dans un contexte bruité, les hautes fréquences semblent donc prépondérantes pour détecter les sons d'eau. Une analyse acoustique des sons d'eau et notamment des descripteurs permettant de les discriminer semble être une approche qui permettrait d'affiner cette stratégie.

La plupart de ces méthodes nécessitent un apprentissage supervisé. Dans ce cas, un sous ensemble du corpus de l'étude permet de fournir les données d'apprentissage qui servent à calculer les paramètres du système. Ce type d'approche donne des résultats très satisfaisants quand les données d'apprentissage et de test sont homogènes. Dans la plupart de ces études, les capteurs utilisés sont placés dans les mêmes lieux pour l'apprentissage et le test, ce qui réduit largement les risques d'erreur. Si Taati fait mention de plusieurs lieux, ils semblent avoir des propriétés similaires, et la caméra est toujours placée au même endroit par rapport au lavabo.

Nous avons déjà largement mentionné les contraintes de notre projet IMMED : la diversité des lieux d'enregistrement, le déplacement de la caméra mobile, et la présence importante de voix. Ainsi, dans un souci de création d'un système robuste, nous avons commencé à sélectionner des descripteurs du signal permettant de discriminer les sons d'eau de façon robuste dans plusieurs situations, et notamment en présence de parole.

Avant de présenter cette analyse des différents descripteurs, nous allons dans un premier temps définir plus précisément ces sons d'eau que nous cherchons à reconnaître.

3.2 Le flux d'eau

3.2.1 Définition

Le système présenté dans ce chapitre n'est pas destiné à reconnaître tous types de sons d'eau, mais principalement les sons d'eau liés à l'utilisation du robinet dans le cadre d'activité de la vie quotidienne. D'autres types de reconnaissance de son d'eau seront présentés dans le chapitre

suivant. Les sons qui nous intéressent dans ce chapitre sont concrètement produits par « de l'eau qui coule ». Ce phénomène est appelé *water flow* dans la littérature. En français nous appelons ce phénomène « flux d'eau », en référence à sa définition :

Flux : « Écoulement d'un liquide organique ou de matière liquide en général », (Petit Larousse).

Le flux a la particularité de décrire un ensemble de particules, ici les particules d'eau, évoluant dans un sens commun, et qui ont une origine, un trajet et une destination similaires. Ce terme semble donc bien adapté pour décrire le phénomène de l'eau qui coule et tombe depuis un robinet.

Par ailleurs, le terme jet est également proche de cette idée, mais suggère l'idée que le fluide jaillit avec force. Par abus de langage, et par analogie avec l'anglais, nous avons régulièrement appelé ce phénomène « flot d'eau ». Le flot, étant un terme généralement employé pour une masse de liquide en mouvement, sans direction précise.

Flot : « Toute masse liquide agitée de mouvements en sens divers » (Petit Larousse).

Nous définissons donc un « flux d'eau » comme la chute de l'eau s'écoulant d'un robinet ouvert. Cette définition peut s'étendre à une douche ou même à une cascade.

3.2.2 Modèle acoustique

Au niveau acoustique nous nous intéressons aux sons produits par ces flux d'eau. La surface sur laquelle tombe l'eau a évidemment beaucoup d'importance dans l'acoustique de ces sons mais ne sera pas prise en compte pour l'instant. Nous supposerons par la suite que la surface sur laquelle l'eau tombe est stable et faiblement résonnante. Un modèle plus précis de l'acoustique des sons d'eau sera présenté dans le chapitre 4.

Nous pouvons identifier quelques propriétés acoustiques des sons de flux d'eau, que nous allons pouvoir observer sur un exemple. La figure 3.7 présente un spectrogramme de son de flux d'eau. Cet enregistrement a été effectué avec un robinet fortement ouvert au dessus d'un lavabo.

Cet exemple n'est évidemment pas du tout représentatif de la grande variété de sons apparaissant dans la vie quotidienne. De plus, dans le contexte dégradé du projet IMMED, les sons de flux d'eau sont en général recouverts par de la parole ou des sons d'impacts, et peuvent varier selon des déplacements d'objets sous le jet d'eau. Cet extrait représente enfin la partie continue du jet, mais l'ouverture et la fermeture créent des perturbations et un changement dans l'enveloppe temporelle.

Néanmoins, cette approximation nous permet de travailler sur un modèle. Ainsi, d'après la figure 3.7, ces sons ont la particularité d'être bruités : la plupart des fréquences, hormis les fréquences les plus graves, semblent présentes dans le spectre. De plus, l'enveloppe temporelle de ce son semble stable, ce qui amène une notion de continuité. Nous allons dans la partie suivante nous attacher à trouver des descripteurs permettant de mettre en évidence l'aspect bruité de ces sons. La continuité sera prise en compte dans notre système de reconnaissance décrit dans la partie 3.4.

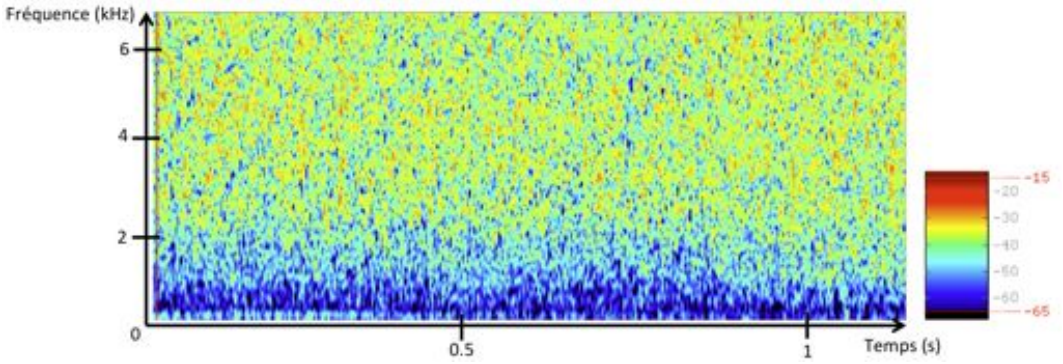


FIGURE 3.7 – Spectrogramme de son de flux d'eau.

3.3 Descripteurs acoustiques

3.3.1 Introduction

Afin de juger de la pertinence des descripteurs pour la détection de flux d'eau, nous avons analysé sur plusieurs exemples les descripteurs classiquement utilisés en reconnaissance sonore. La figure 3.8 représente ainsi un signal de 5 minutes extrait du corpus IMMED. Une partie de ce signal sonore a été enregistrée lors de l'activité *faire la vaisselle*. Sur le spectrogramme, nous pouvons observer trois minutes bruitées et continues, qui correspondent au flux d'eau du robinet ouvert. De plus, nous pouvons apercevoir à la fin des 3 minutes de flux d'eau que le signal est particulièrement bruité. Ce passage correspond à un temps de dialogue où la parole se superpose au flux d'eau.

En dehors de ce flux d'eau, au début et à la fin du signal, il semble que l'ambiance ne soit pas non plus silencieuse. Les sons de dialogue entre le patient et l'assistant médical, ainsi que divers bruits de la maison et de l'environnement extérieur, coexistent dans ces parties.

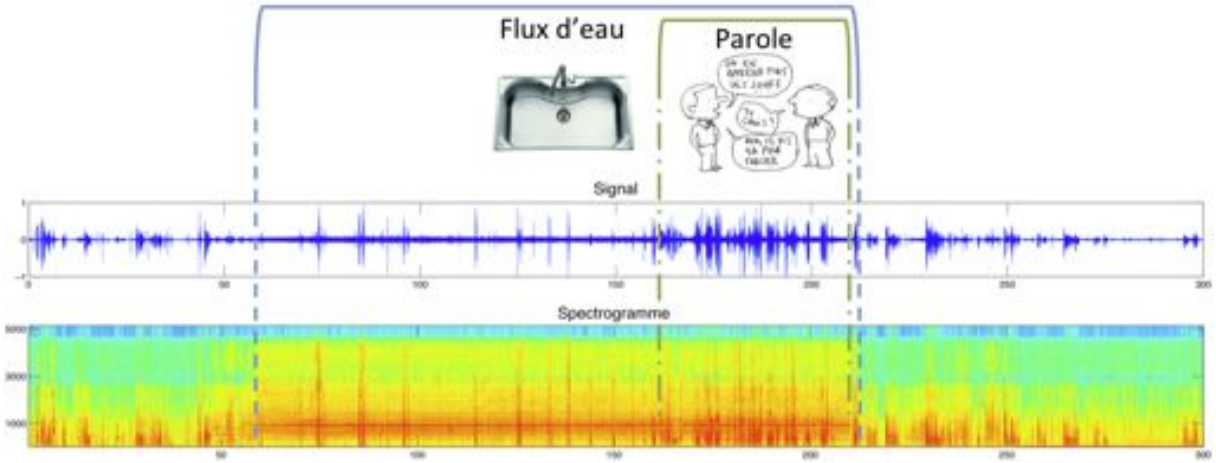


FIGURE 3.8 – Spectrogramme d'un son de flux d'eau lors de l'activité *faire la vaisselle*.

3.3.2 Descripteurs usuels

De nombreux descripteurs usuels ont été calculés sur cet extrait. Nous avons de plus étendu notre choix de paramètres à des descripteurs moins utilisés, mais connus pour quantifier l'aspect bruité d'un signal, par exemple le coefficient de variation [WW94] et le *spectral flatness* [Joh88].

Nous avons ainsi calculé des paramètres :

- temporels comme *l'énergie, l'énergie par bande, et le ZCR*,
- et spectraux : *flux, centroïde, spread, skewness, kurtosis, flatness, roll-off à 95%, coefficient de variation.*

Les formules de ces paramètres sont rappelées en annexe C.

La figure 3.9 nous montre les courbes de variation de trois descripteurs de cet ensemble. Ces trois descripteurs sont extraits du signal et calculés sur des fenêtres de hamming de 80 ms, avec 40 ms de recouvrement. Pour des facilités de lecture, ces courbes sont lissées par un filtre médian sur une fenêtre de 4 secondes. D'autres types de fenêtrage et de filtrage ont également été essayés dans cette étude.

Nous pouvons voir sur la figure 3.9 les courbes de ces descripteurs, dont celle du *Zero Crossing Rate*. La courbe (c) du ZCR augmente sensiblement en présence du flux d'eau par rapport au reste de l'enregistrement. Ceci peut s'expliquer par le fait que le flux d'eau amène des fréquences aiguës.

Par contre, la courbe (c) baisse en présence de parole superposée au flux d'eau. La parole présente ainsi beaucoup d'énergie dans les fréquences graves : le formant F_1 se situe en général en dessous de 1000 Hz. Les fréquences graves de la voix peuvent diminuer le ZCR. Rappelons que la voix peut avoir une énergie importante par rapport au son du flux d'eau, le microphone étant proche de la bouche du patient.

La courbe (d) présente l'évolution du centroïde spectral. Nous pouvons voir que cette courbe est assez proche de celle du ZCR sur cet exemple. Le centroïde spectral augmente également en présence de flux d'eau mais oscille et diminue lorsque la parole se superpose à celui-ci.

La courbe (e) montre l'évolution du *spectral flatness*, régulièrement employé pour quantifier l'aspect bruité d'un son. Ce descripteur semble un peu moins sensible à la présence de la parole superposée au flux d'eau. Toutefois, en dehors du flux d'eau, il est très sensible à l'aspect bruité du signal, même pour les bruits de faible énergie.

3.3.3 Un nouveau descripteur : la couverture spectrale

Nos recherches nous ont amenés à travailler sur l'élaboration d'un nouveau descripteur, capable de réagir à l'aspect bruité du son tout en restant relativement constant en présence de parole. Idéalement, ce descripteur ne devrait pas être trop sensible aux bruits en dehors du flux d'eau, et ne devrait pas varier en fonction de l'énergie du signal.

Nous avons ainsi créé un nouveau descripteur, que nous avons appelé *spectral cover*, ou couverture spectrale. Le terme *couverture* fait référence à l'importante bande de fréquence présentant de l'énergie dans un signal bruité. La couverture spectrale est définie ainsi :

$$SpectralCover = \frac{\sum_{n=1}^{N} [f(k)w(k)]^2}{[\sum_{n=1}^{N} w(k)]^\gamma} \tag{3.1}$$

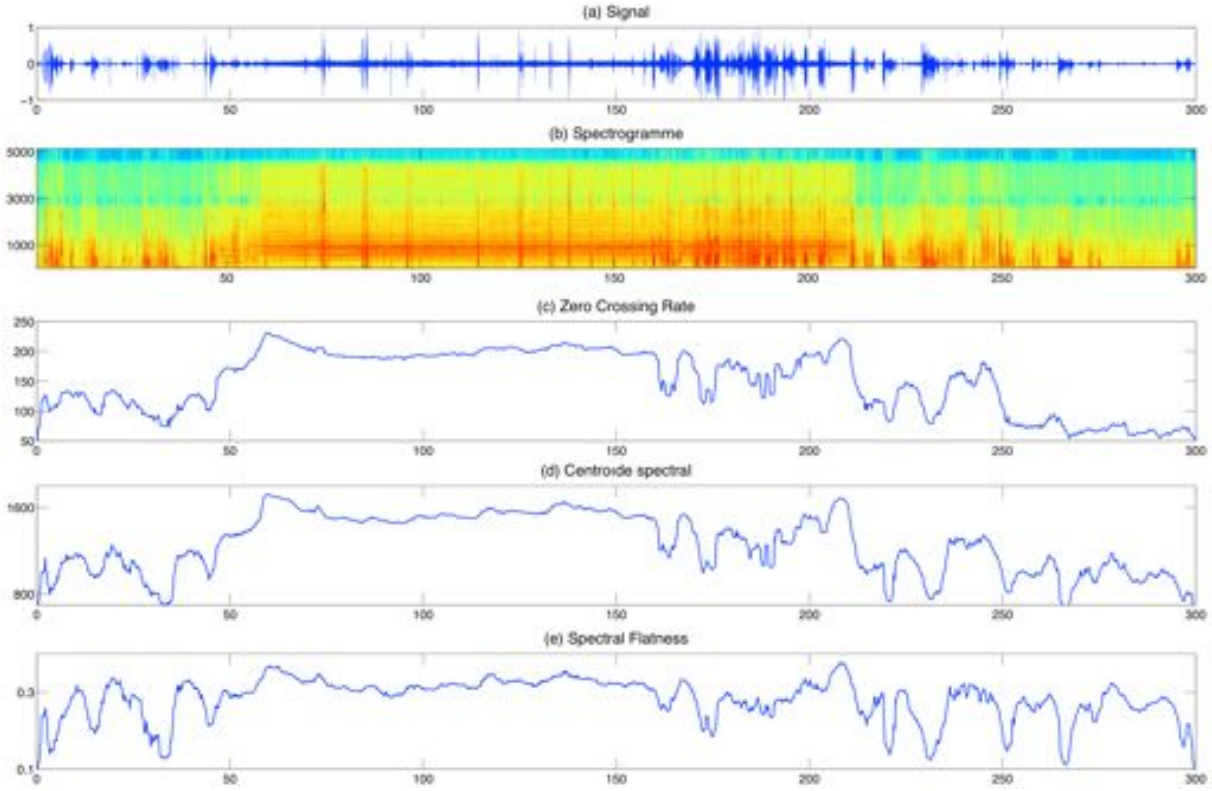


FIGURE 3.9 – Descripteurs calculés sur l'activité *faire la vaisselle*.

où $w(k)$ est l'amplitude de la fréquence $f(k)$. Pour comparaison, voici un rappel de l'équation du centroïde spectral :

$$CSG = \frac{\sum_{n=1}^{N} f(k)w(k)}{\sum_{n=1}^{N} w(k)} \tag{3.2}$$

Nous faisons intervenir un carré au numérateur de la couverture spectrale qui va amplifier les hautes fréquences. Le paramètre γ permet de faire varier ce descripteur en fonction de l'énergie totale du signal, à la différence par exemple du centroïde spectral.

La figure 3.10 montre la courbe de la couverture spectrale, calculée sur le même extrait que précédemment. Le paramètre γ est fixé à $\gamma = \frac{3}{2}$. Le même lissage temporel a été appliqué.

Nous observons une augmentation de la couverture spectrale avec l'apparition du flux d'eau. Contrairement aux descripteurs précédents, elle reste relativement stable et à un niveau élevé en présence de parole. Sur cet exemple, il semble ainsi relativement facile d'identifier la partie liée au flux d'eau par l'intermédiaire de la couverture spectrale. Un simple seuil semble suffire à identifier deux zones, comme nous le voyons sur la figure 3.11.

3.3.4 Conclusion

Afin de reconnaître automatiquement le flux d'eau, nous avons analysé le comportement d'un ensemble de descripteurs usuels sur des extraits du corpus IMMED. Pour les descripteurs usuels

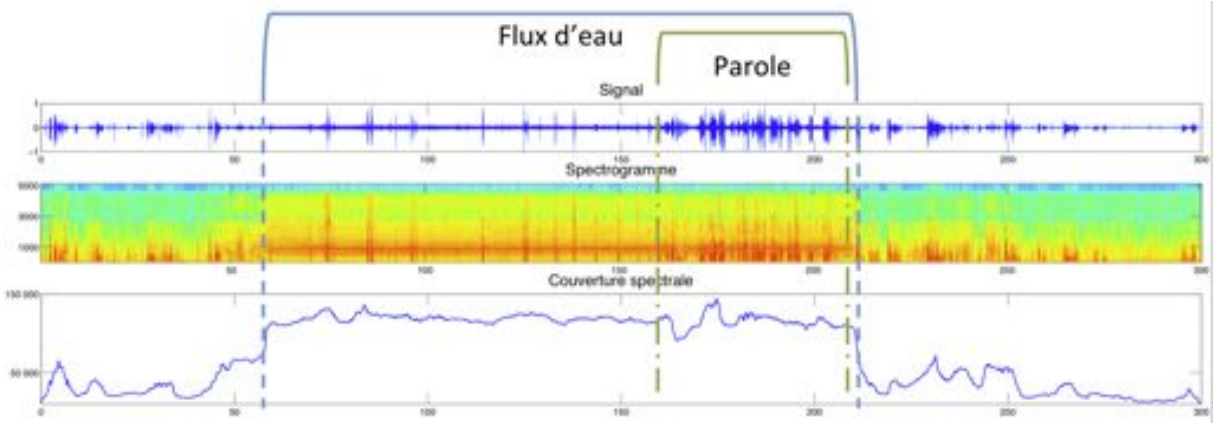


FIGURE 3.10 – Couverture spectrale calculée sur l'activité *faire la vaisselle*.

FIGURE 3.11 – Seuillage sur la couverture spectrale pour déterminer la zone de flux d'eau.

bien adaptés à quantifier l'aspect bruité d'un signal, la présence de parole est problématique. La forte présence d'harmoniques graves les fait varier de manière importante.

Un nouveau descripteur, la couverture spectrale, nous permet d'identifier correctement un flux d'eau à l'aide d'un simple seuil, malgré la présence de parole superposée au flux d'eau. Il faut maintenant valider notre approche, en particulier la faisabilité d'un simple seuillage, sur l'ensemble du corpus IMMED qui comporte des sons très hétérogènes, et . Dans la partie suivante, nous allons décrire notre système de reconnaissance de flux d'eau.

3.4 Système de reconnaissance de flux d'eau

Nous proposons d'explorer un système de reconnaissance de flux d'eau basé sur la couverture spectrale et d'un seuil T_1. La description de ce système et les expériences effectuées dans cette partie ont été publiées dans la conférence *Content-Based Multimedia Indexing* [GPAO12].

3.4.1 Au delà d'un simple seuillage

Le calcul de la couverture spectrale sur un fichier de 39 minutes issu du projet IMMED nous a permis d'observer quelques difficultés et d'affiner notre système. La figure 3.12 montre une analyse de 4 minutes issue du logiciel *Sonic Visualizer*[13]. Nous pouvons observer la forme temporelle du signal en haut en noir, et les valeurs de couverture spectrale en bas en rouge. La couverture spectrale n'est ici pas lissée, contrairement à la section précédente. Ceci explique les

13. http://www.sonicvisualizer.org

grandes discontinuités, qui sont produites par des bruits divers dans l'enregistrement. L'écoute du fichier nous permet d'identifier un aspirateur, un bruit causé par des déplacements d'objets lors d'une recherche dans un tiroir, et trois sons de flux d'eau.

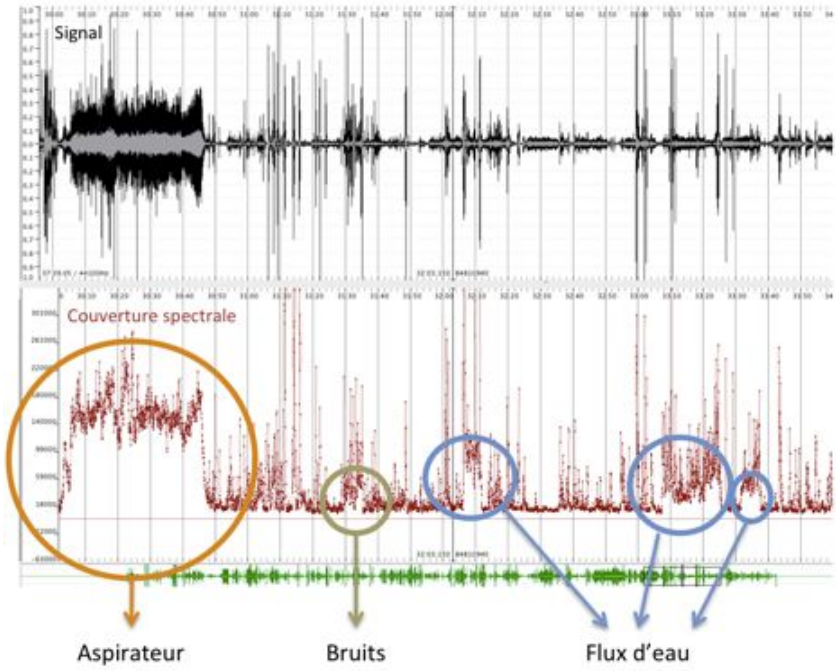


FIGURE 3.12 – Couverture spectrale sur un extrait de 4 minutes du corpus IMMED.

Notre analyse nous a permis de dresser les conclusions suivantes.

- Les sons d'aspirateurs, présents dans l'activité *faire le ménage*, provoquent des valeurs de couverture spectrale très élevées. Pour éviter les fausses alarmes, nous allons donc utiliser un deuxième seuil T_2.
- D'autres sons créent également des fausses alarmes, tels le chant des oiseaux, le frottement de sacs plastiques et certains segments de parole. Pour être robuste à ces sons, nous allons utiliser l'aspect continu des flux d'eau en étudiant le signal sur des fenêtres temporelles de 2 secondes.

Nous allons vérifier que ce choix nous permet d'être robuste à la plupart des fausses alarmes créées par des sons forts et aigus de durée inférieure à 2 secondes. En contrepartie, avec une fenêtre d'une durée importante, nous contraignons notre système à la détection de sons d'une durée supérieure à deux secondes. Cette opération constitue néanmoins un compromis réaliste sur le corpus IMMED ou la grande majorité des sons de flux d'eau présentent une durée supérieure.

3.4.2 Présentation du système

La figure 3.13 illustre notre système de reconnaissance sous forme de diagramme. Nous allons maintenant détailler chaque étape de ce système.

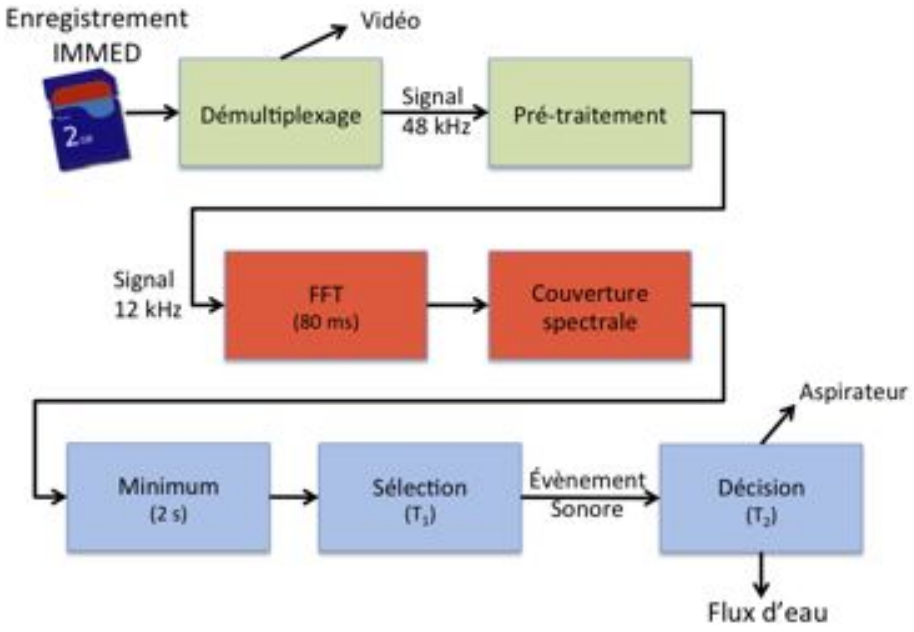


FIGURE 3.13 – Diagramme de notre système de reconnaissance.

Démultiplexage

Les vidéos enregistrées au domicile du patient par la caméra GoPro (voir chapitre 1) sont stockées sur une carte SD. L'opération de démultiplexage consiste à séparer l'audio de la vidéo sur ces fichiers. Cette opération est réalisée grâce à la bibliothèque *ffmpeg*[14]. À l'issue de cette opération, le signal sonore stéréo est échantillonné à 48 kHz.

Pré-traitement

Dans ce signal, les deux canaux sont identiques. Nous ne conservons donc qu'un seul des deux canaux. De plus, comme nous l'avons expliqué au chapitre 1, les fréquences les plus élevées résultent d'un repliement spectral. Nous supprimons ces fréquences, en ne conservant que la partie du signal située en dessous de 6 kHz.

Le filtrage des hautes fréquences est effectué sur une transformée de Fourier à l'aide du programme *supervp*[15], noyau du logiciel *Audiosculpt*[16]. Les coefficients de la transformée de Fourier correspondant aux hautes fréquences sont mis à zéro, et le signal est reconstruit par le programme *supervp* basé sur le vocodeur de phase. Cette méthode originale permet de s'affranchir de la pente d'un filtrage passe-bas en « coupant » les fréquences de manière précise. À l'issue de cette opération, le signal est sous-échantillonné à 12 kHz.

FFT

Suite à un fenêtrage de Hamming sur des trames de 1024 points (environ 85 ms), nous calculons la transformée de Fourier par l'algorithme *Fast Fourier Transform*. Le recouvrement

14. http://www.ffmpeg.org/
15. http://anasynth.ircam.fr/home/software/supervp
16. http://anasynth.ircam.fr/home/software/audiosculpt

pour chaque analyse est de moitié. Nous utilisons des fenêtres relativement larges par rapport à la plupart des travaux de reconnaissance sonore, car nous privilégions la précision fréquentielle à la précision temporelle. Une fenêtre de 8192 échantillons, soit 185 ms, avait été utilisée dans [VSN+11] pour permettre une sélection précise des coefficients de Fourier.

Couverture spectrale

Nous calculons ensuite la couverture spectrale, grâce à la formule suivante :

$$SpectralCover = \frac{\sum_{n=1}^{N} [f(k)w(k)]^2}{[\sum_{n=1}^{N} w(k)]^{\frac{3}{2}}} \tag{3.3}$$

Minimum

Sur chaque fenêtre de 2 secondes, nous calculons le minimum des coefficients de la couverture spectrale. Les fenêtres sont décalées d'une trame à chaque analyse.

Sélection

Cette courbe des valeurs minimales de la couverture spectrale est seuillée. Tout segment constitué de valeur supérieur au seuil T_1 est considéré comme un évènement sonore.

Décision

Nous considérons chaque segment d'évènement sonore. Si les valeurs de couverture spectrale sont supérieures à un seuil T_2, ce segment est considéré comme un son d'aspirateur. Sinon il est considéré comme un son de flux d'eau. Le résultat de la décision sur notre exemple est représenté sur la figure 3.14.

3.4.3 Expériences

Corpus

Les expériences ont été effectuées sur un corpus de 20 vidéos issues du projet IMMED. Chaque vidéo est associée à un patient différent, qui effectue des activités de la vie quotidienne à son domicile. Ce corpus a donc été enregistré dans 20 lieux de vie différents.

Les vidéos ont été annotées en activités de façon manuelle par l'assistant médical du projet IMMED. Une annotation supplémentaire a été nécessaire pour définir précisément les bornes de chaque son d'eau. Pour des raisons d'efficacité et de rapidité, nous nous sommes appuyés sur la vidéo pour réaliser cette annotation. De ce fait, les segments dans lesquels un robinet ouvert apparaît à l'image ont été annotés en sons d'eau même lorsque le son d'eau est difficilement audible, par exemple quand il est recouvert par d'autres sons. Une annotation uniquement basée sur le son nécessiterait plusieurs annotateurs pour être plus objective et se révèlerait très coûteuse. De plus, l'annotation d'un fichier est dans tous les cas influencée par la compréhension du contexte.

Cette annotation est d'autant plus délicate que les sons d'eau montrent une grande variabilité acoustique. L'annotation précise en « flux d'eau » est donc techniquement très difficile devant l'ensemble des sons produits par l'eau. Le corpus a donc été annoté en son d'eau, et non en flux.

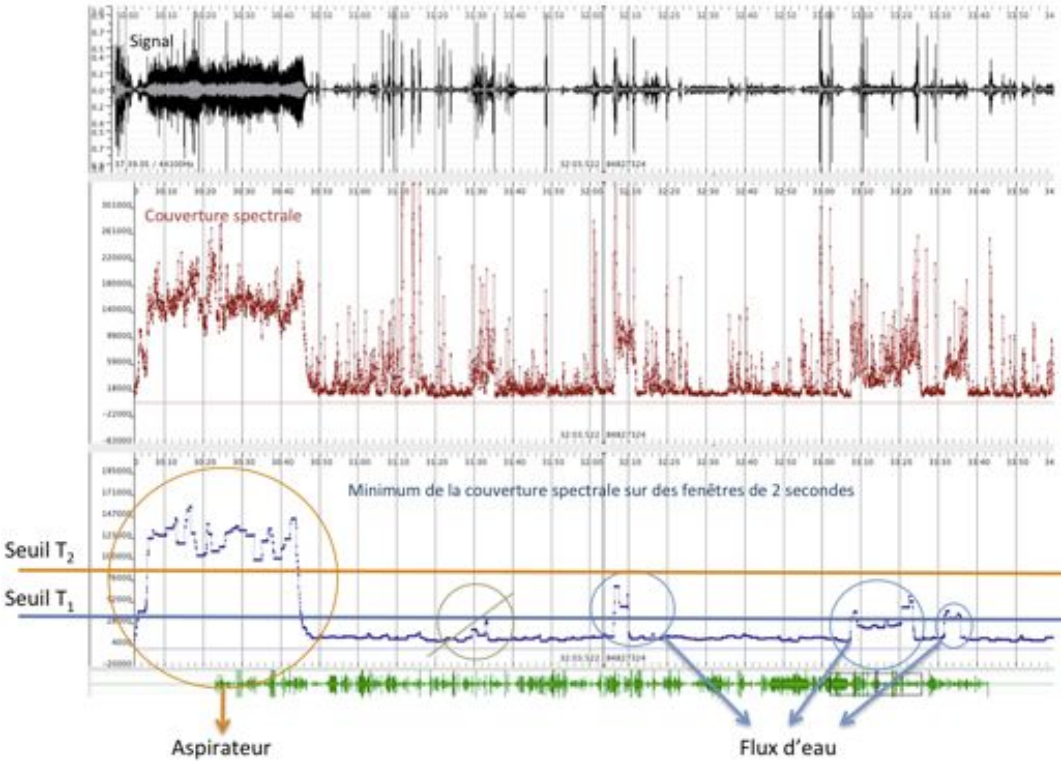


FIGURE 3.14 – Illustration des seuils sur un extrait de 4 minutes.

Ainsi, bien que nous distinguons plusieurs types de sons d'eau, nous utilisons dans les expériences une annotation unique en son d'eau.

Le corpus a également été annoté en sons d'aspirateurs. Cette annotation s'est révélée beaucoup plus simple que celle des sons d'eau, la seule ambiguïté de l'annotation résidant éventuellement dans la détermination précise des frontières du début et de la fin de l'évènement sonore. Au final, le corpus présente les caractéristiques suivantes :

- durée totale : 7 heures et 30 minutes,
- durée minimale d'une vidéo : 6 minutes,
- durée maximale d'une vidéo : 44 minutes,
- nombre total de sons d'eau : 85,
- durée totale des sons d'eau : 25 minutes,
- nombre total des sons d'aspirateurs : 9,
- durée totale des sons d'aspirateurs : 10 minutes.

Développement

Une vidéo de 39 minutes a été utilisée pour le développement. Sur ce fichier, 13 sons d'eau sont présents, pour une durée totale de 3 minutes. Deux sons d'aspirateurs sont également présents. Nous avons fixé les valeurs des seuils sur cette vidéo :

$$T_1 = 30000 \qquad et \qquad T_2 = 80000 \tag{3.4}$$

De plus, pour être robuste aux démarrages et aux arrêt de l'aspirateur, nous avons légèrement modifié notre étape de décision. Le segment d'évènement sonore est considéré comme un son d'aspirateur si 85% des valeurs de la couverture spectrale sont supérieures au seuil T_2.

Enfin, dans cette expérience, nous appliquons un post-traitement supplémentaire et les sons de flux d'eau inférieurs à 3 secondes sont supprimés.

Résultats

Nous avons considéré deux types d'évaluation :
- temporelle, pour apprécier les durées de documents correctement étiquetés,
- en évènements sonores (voir chapitre 2), pour estimer la qualité du lecteur à pointer chaque segment de son d'eau.

Néanmoins, les scores de ces deux évaluations étant relativement similaires, nous avons choisi par la suite d'utiliser uniquement les métriques temporelles utilisées dans la campagne d'évaluation ESTER2 [GGM+05] :
- taux d'erreur,
- précision,
- rappel,
- F-mesure.

Aspirateur Les résultats de la reconnaissance des sons d'aspirateur sont très bons, avec une F-mesure de près de 99%. L'étude acoustique de ces appareils pourrait faire l'objet d'une recherche plus approfondie, mais nous nous sommes focalisés dans cette thèse sur la reconnaissance des sons d'eau, car celle-ci a notamment le mérite de représenter un nombre plus important d'activités du quotidien.

Flux d'eau Les résultats de la reconnaissance de sons d'eau obtiennent une F-mesure de 66% (voir tableau 3.1).

TABLE 3.1 – Tableau des résultats de notre système.

	Taux d'erreur	Précision	Rappel	**F-mesure**
Notre système	**5%**	**54%**	83%	**66%**

Parmi les évènements manqués, certains sons d'eau sont difficilement audibles. La vérité terrain ayant été réalisée grâce à la vidéo, ceci questionne sur les capacités d'un humain à entendre et reconnaître la totalité des segments annotés comme sons d'eau.

De plus, certains sons annotés comme son d'eau ne comportent pas, ou peu de flux d'eau. Ils correspondent par exemple à des gouttes tombant dans un évier rempli, ou à des mouvements d'eau. Notre système ne peut reconnaître ce type de son. Nous verrons par la suite que ces sons sont pourtant immédiatement identifiables par un humain comme des sons d'eau.

Le système a également produit 22 fausses alarmes. 6 erreurs proviennent de la manipulation d'objets tels que les sacs en plastiques. D'autres sont dues à la télévision (3 erreurs) ou à des manipulations de la caméra (2 erreurs). D'autres encore viennent de bruits vocaux, comme les rires (4 erreurs). Enfin la parole superposée provoque 7 erreurs dont 6 dans la même session d'enregistrement (dans laquelle le patient parle très fort).

3.4.4 Comparaison avec un système classique

Pour justifier de l'utilité de notre système par rapport aux systèmes classiquement utilisés dans l'état de l'art, nous avons effectué une étude comparative.

Paramétrisation

Plusieurs ensembles de descripteurs ont été calculés et testés. Au final, nous présentons dans cette expérience les ensembles suivants :

- **MFCC** (Mel Frequency Cepstral Coefficients) : 24 coefficients cepstraux représentant des fréquences allant de 20 Hz à 6 kHz.
- **LLD** (Low Level Descriptors) : *énergie, centroide spectral*, *spectral spread*, *spectral skewness* et *spectral kurtosis*. Nous ajoutons également le *spectral flatness*, calculé par bande de fréquences espacées de $\frac{1}{4}$ d'octave, tel que présenté dans le standart MPEG7 [KMS06]. Ces descripteurs ont été calculés à l'aide de la bibliothèque *Yaafe audio extractor*[17].
- **LLD+CS** : notre paramètre « couverture spectrale » est ajouté au système LLD.

Classification

Nous avons utilisé un classifieur GMM, régulièrement utilisé comme référence dans l'état de l'art. Nous avons testé ce système avec différents nombres de lois gaussiennes (2, 4, 8, 16). Au final le GMM de 8 gaussiennes produit les meilleurs résultats.

Par ailleurs, un post-traitement similaire à celui de notre système est utilisé. Les segments de flux d'eau inférieurs à 3 secondes sont supprimés. Nous appliquons une approche *leave-one-out* sur cette classification : 19 vidéos sont utilisées en apprentissage et 1 vidéo en test. Comme les vidéos représentent 20 lieux et 20 patients, les ensembles de test et d'apprentissage correspondent à des lieux et des patients différents pour chaque itération.

Résultats

Les résultats sont présentés dans le tableau 3.2. Les scores sont globalement assez inférieurs à ceux de notre système. Les MFCC obtiennent un score de 45%. L'ensemble de descripteurs LLD obtient ici un score supérieur aux MFCC. Ce résultat pourrait alimenter les discussions autour du choix des descripteurs pour la reconnaissance de sons environnementaux [KS04]. Il est intéressant de constater que l'ajout de la couverture spectrale à l'ensemble des LLD améliore le résultat. Cette expérience confirme l'intérêt de ce descripteur qui semble contenir une information complémentaire de celle des descripteurs de l'ensemble LLD.

Table 3.2 – Tableau comparatif des systèmes de reconnaissance de flux d'eau.

	Taux d'erreur	Précision	Rappel	**F-mesure**
GMM / MFCC	9%	35%	87%	45%
GMM / LLD	8%	39%	91%	50%
GMM / LLD-CS	7%	44%	88%	53%
Notre système	**5%**	**54%**	83%	**66%**

17. http://yaafe.sourceforge.net/

3.4.5 Conclusion

Pour reconnaître les flux d'eau, nous avons élaboré un système assez simple basé sur un seuillage de notre descripteur « couverture spectrale ». Ce système, appliqué à 20 vidéos du corpus IMMED enregistrées dans des lieux différents, permet d'obtenir pour la tâche de reconnaissance de flux d'eau une F-mesure de 66%. Sur ce même corpus, les approches classiques obtiennent des résultats bien inférieurs. Ces scores confirment la pertinence de notre descripteur, la couverture spectrale, qui est adaptée aux caractéristiques acoustiques du son à reconnaître. Les résultats montrent également la difficulté de l'utilisation d'un système nécessitant de l'apprentissage sur un corpus hétérogène.

Plus précisément, les scores indiquent globalement une mauvaise précision des systèmes. Notre système affiche une précision de 54%, ce qui signifie que de nombreuses fausses alarmes sont générées. Nous proposons dans une étape la partie 3.6 une méthode pour réduire le nombre de ces fausses alarmes, en rajoutant un traitement supplémentaire à notre système.

Par ailleurs notre système affiche de très bons résultats pour la reconnaissance de sons d'aspirateur. Au final, notre système permet de détecter l'utilisation de l'eau et de l'aspirateur de façon satisfaisante pour être utilisé dans le cadre plus général de l'identification d'activités. Nous l'avons donc intégré aux outils développés pour le projet IMMED. Afin de contribuer à la reconaissance d'activités, les résultats de nos segmentations en sons d'eau et en sons d'aspirateurs ont été associés à des paramètres issus de la vidéo et à d'autres paramètres sonores au sein d'un modèle audio-vidéo.

3.5 Intégration au projet IMMED

3.5.1 Modèle d'activités

Au sein du projet IMMED, la reconnaissance automatique des activités à partir de l'audio et de la vidéo est basée sur un Modèle de Markov Caché Hiérarchique (en anglais HHMM), développé par le LaBRI dans le cadre du projet [Kar11]. La figure 3.15 illustre ce HHMM à deux niveaux. Le premier niveau représente l'enchaînement des activités effectuées par le patient. L'activité Act_i est ici développée au niveau inférieur par un nombre d'états non-sémantiques, ici fixé à trois.

3.5.2 Paramètres du modèle d'activités

Les paramètres de ce modèle, extraits de la vidéo et de l'audio, sont regroupés en trois modalités : mobilité, localisation, évènements sonores. L'analyse des caractéristiques du mouvement dans la vidéo produit le premier ensemble : les paramètres visuels dynamiques [Kar11]. Le deuxième ensemble est créé à partir de paramètres visuels statiques, dont les paramètres de couleur et de localisation [Dov11]. Enfin, le troisième ensemble est constitué des paramètres audio.

3.5.3 Paramètres audio

Nous avons associé notre système de détection de flux d'eau et de sons d'aspirateur aux outils de l'IRIT déjà intégrés dans le projet. Ces outils, qui ont été développés en amont du projet sur des corpus radiophoniques, permettent d'obtenir des indices sur la présence de parole, de musique, ou de bruit dans un signal sonore [Pin04].

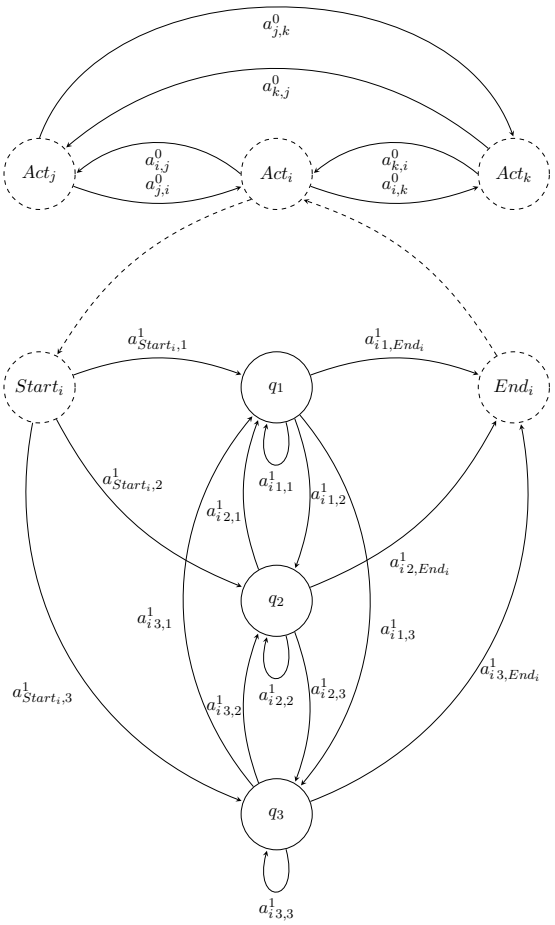


FIGURE 3.15 – Modèle de Markov Caché Hiérarchique pour la détection des activités [KBPD+11].

Techniquement, la modulation d'entropie et la modulation d'énergie à 4 Hz donnent un indice sur la présence de parole. Le nombre et la durée des segments issus d'une segmentation du signal [AO88] fournissent un indice sur la présence de musique. L'énergie permet d'inférer sur le silence. Des seuils sont utilisés sur ces paramètres pour obtenir les probabilités. De plus, Un système GMM-MFCC entrainé sur le corpus Ester2 [GGM+05], est utilisé dans le but de détecter des sons percussifs et périodiques. Au final, en ajoutant notre système de détection d'eau et d'aspirateur, nous obtenons 7 types d'évènements sonores, illustrés sur la figure 3.16.

Le découpage de la vidéo par rapport au mouvement de la caméra [KBPM+10] permet d'obtenir des segments sur lesquels nous calculons nos paramètres. Pour chaque type d'évènements sonores, une probabilité par segment est calculée. Nous avons adapté nos systèmes de détection de son de flux d'eau et de son d'aspirateurs, pour obtenir un ensemble de paramètres homogènes. La probabilité des sons d'eau et des sons d'aspirateur est calculée respectivement en fonction de la proportion d'eau et d'aspirateur détectée sur chaque segment du découpage vidéo.

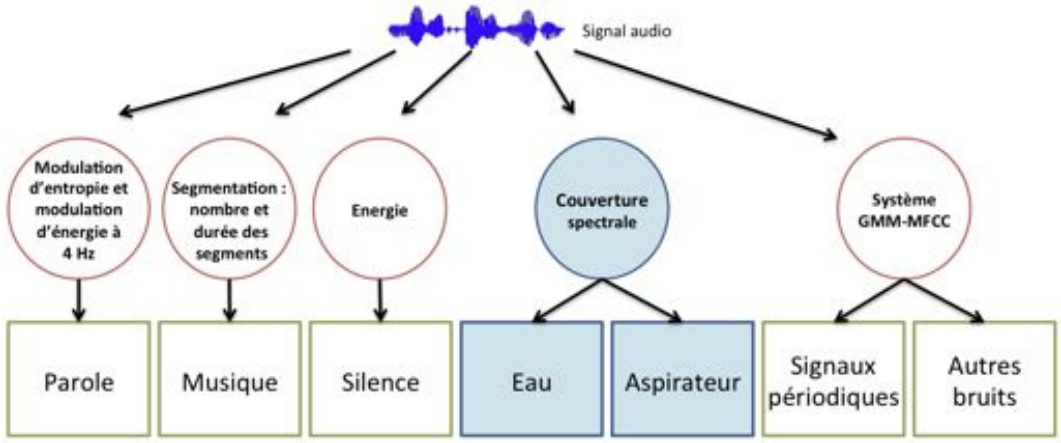


FIGURE 3.16 – Système d'extraction de paramètres audio.

3.5.4 Fusion de flux audio/vidéo

La fusion des paramètres issus des différentes modalités (mobilité, localisation, évènements sonores) dans le HHMM a fait l'objet d'un étude particulière [PKL+12]. En effet, selon l'activité à reconnaître, les informations de mouvement, de localisation, ou des bruits caractéristiques ne possèdent pas le même intérêt et peuvent apporter plus ou moins d'information. L'espace des paramètres est ainsi divisé en trois sous-espaces :

- les paramètres visuels dynamiques de mouvement de dimension 34,
- les paramètres visuels statiques de dimension 19, comprenant les sous-espaces des paramètres de couleur et de localisation,
- les paramètres audio de dimension 7.

Plusieurs types de fusion ont été proposés :

- une fusion précoce, où tous les descripteurs, qu'ils soient audio ou vidéo ont le même poids. Dans cette approche, une observation du HHMM correspond à la concaténation des trois modalités (mouvement, localisation, évènements sonores) dans un espace de dimension 60.
- une fusion intermédiaire où une observation est constituée d'un ensemble d'observations de cardinalité égale au nombre de modalités, ici 3. Chaque modalité a ainsi le même poids.
- une fusion tardive où chaque modalité donne lieu à une prise de décision sur l'activité reconnue. Ce modèle prend ainsi en compte le fait que les modalités puissent apporter une information plus ou moins pertinente selon le type d'activité à reconnaître. La fusion est alors réalisée sur les différentes décisions prises indépendamment.

Sur les expériences effectuées, la fusion intermédiaire donne les meilleurs résultats, avec 25% de F-mesure dans la tâche d'indentification d'activités pour 24 activités à identifier [PKL+12].

3.5.5 Conclusion

Notre système de détection de flux d'eau et de son d'aspirateur a été intégré au projet IMMED. Il complète les outils déjà utilisés dans le projet, et permet de contribuer à la segmentation des vidéos en activités. Toutefois, si les sons d'aspirateurs sont facilement détectés dans ce corpus, la détection des sons d'eau n'est pas optimale et fait apparaître des erreurs. Dans la perspective de fournir au modèle de reconnaissance d'activités une segmentation en sons d'eau la plus précise possible, nous nous sommes concentrés sur l'amélioration de notre système.

3.6 Amélioration du système par une étape de classification

Notre approche de détection de flux d'eau a été enrichie suite à une étude effectuée en collaboration avec Xavier Valéro, doctorant encadré par Francesc Alias à l'Université Raymond-Lulle de Barcelone. Cette étude a donné lieu à une publication dans la conférence *International Conference on Multimédia and Expo* [GVPA13].

3.6.1 Présentation du système

La méthode développée vise à améliorer notre système de détection de flux d'eau, fondé sur la couverture spectrale, par l'ajout d'une étape de classification. En effet, selon la précision de notre système (54% dans la section 3.4), une proportion importante des segments résultats ne correspond pas au son d'eau. Une étape de classification, basée sur un modèle de son d'eau appris, a pour but la suppression des fausses alarmes afin de conserver uniquement les segments de flux d'eau.

Nous appellerons « système initial » le système décrit dans la section 3.4. Les segments produits par ce système initial sont appelés « segments cibles ». Dans notre nouveau système, les segments cibles sont découpés en sous-segments d'une seconde. Ces sous-segments sont soumis à une étape de classification, qui permet de les annoter selon les labels « eau » ou « autre ». Les segments contigus de label « eau » sont ensuite regroupés en segments de flux d'eau. La figure 3.17 montre un diagramme du système global, que nous appelons « système hiérarchique ». Celui-ci est découpé en deux étapes : la segmentation et la classification.

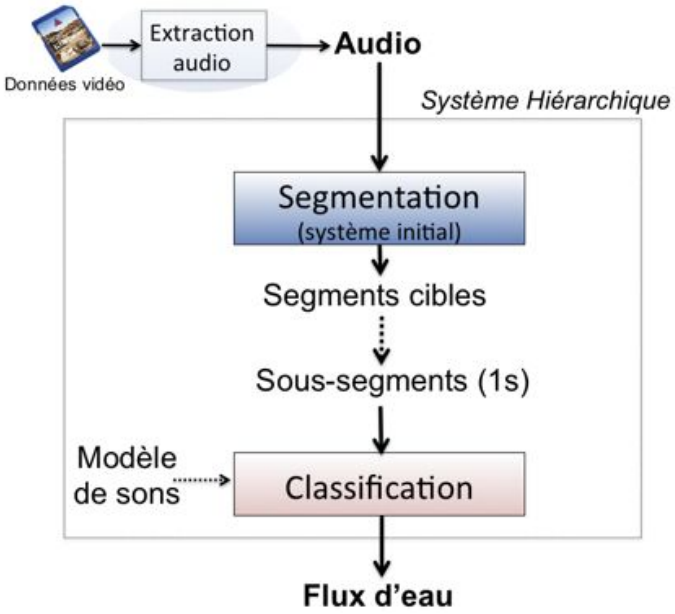


FIGURE 3.17 – Système hiérarchique.

Étape de segmentation

Dans ce système hiérarchique, l'extraction du signal audio et l'étape de segmentation effectuée par le système initial sont analogues à celles décrites dans la section 3.4. Par contre, le

développement de ce système nous a amenés à reconsidérer les seuils utilisés (voir partie 3.6.2).

Étape de classification

Descripteurs Le système de classification emploie des coefficients gammatones, décrits en Annexe C. Les gammatones sont calculés sur des trames de 30 ms, toutes les 15 ms. Le filtrage est adapté à la bande de fréquence et 40 coefficients sont calculés entre 20 Hz et 6 kHz. Après application de la transformée en cosinus discrète, 13 coefficients sont retenus.

Les coefficients gammatones sont extraits des trames d'analyse pour chaque sous-segment d'une seconde. Pour chacun de ces sous-segments, nous calculons la moyenne et l'écart-type des coefficients gammatones. Nous obtenons un vecteur de 26 coefficients par sous-segments.

Classifieur Pour l'étape de classification, les sous-segments sont regroupés par vidéo afin de constituer des ensembles d'apprentissage et de test hétérogènes. 18 vidéos sont ensuite utilisées en apprentissage et 2 en test. Pour chaque sous-segment de l'ensemble d'apprentissage, nous attribuons un label « eau » ou « autre » en fonction de la vérité terrain. Les sous-segments présentant deux annotations différentes dans la vérité terrain sont annotés en fonction de l'annotation majoritaire.

Ce processus est répété dix fois, de manière à tester toutes les vidéos. Trois types de modèles de décision ont été testés dans cette étude : GMM, SVM et kNN.

3.6.2 Mise en œuvre

Segmentation

Le but de l'étape de classification du système hiérarchique est de supprimer les fausses alarmes du système de segmentation. Ainsi, l'étape de classification ne peut trouver un segment de flux d'eau qui n'aurait pas été détecté par celle de segmentation. En d'autres termes le rappel du système hiérarchique ne peut être supérieur à celui du système initial.

Nous avons donc modifié les paramètres du système initial pour améliorer le rappel. Cette modification entraîne un perte de précision (et donc de F-mesure) qui sera compensée par l'étape de classification. Pour améliorer le rappel, nous avons simplement baissé notre seuil T_1, en passant de $T_1 = 30000$ à $T_1 = 25000$. Ce nouveau seuil a été choisi sur l'ensemble de nos fichiers de manière à améliorer le rappel de notre système tout en conservant une F-mesure importante. Cette modification améliore sensiblement le rappel qui passe de 83% à 93%. Les scores de ce système modifié, appelé « système BestRecall » sont présentés dans le tableau 3.3.

Table 3.3 – Comparaison des systèmes (initial et BestRecall).

	Precision	Rappel	F-mesure
Système initial	54 %	**83 %**	66 %
Système BestRecall	28 %	**93 %**	44 %

Classification

Une série d'expériences a été réalisée pour déterminer le modèle de décision le mieux adapté à notre problématique. Comme dans le système hiérarchique, les classifieurs sont appris à partir

des sous-segments. Par contre, à la différence des expériences effectuées sur le corpus IMMED, nous allons tester nos modèles de décision sur un corpus constitué uniquement des sous-segments.

Les ensembles d'apprentissage et de tests sont équivalents à ceux définis dans notre système hiérarchique. Nous utilisons donc les segments cibles issus de 18 vidéos en apprentissage, et les segments cibles issus de 2 vidéos en test, en répétant ce processus 10 fois.

Pour chacun des modèles de décision, plusieurs variantes ont été testées. Nous avons ainsi fait varier le nombre N de gaussiennes pour le modèle GMM, le nombre M de voisins pour le modèle kNN (qui utilise la distance euclidienne), et le type de noyau pour les SVM (dont le noyau linéaire, polynomial, et radial gaussien). Au final, les meilleurs résultats sont obtenus pour un système GMM basé sur $N = 10$ gaussiennes, pour un algorithme kNN avec $M = 3$ voisins, et pour un modèle SVM qui s'appuie sur un noyau radial gaussien. Le tableau 3.4 présente les scores de ces différents modèles de décision.

TABLE 3.4 – Classification des segments cibles.

	Precision	Rappel	F-mesure
SVM	89 %	88 %	**88 %**
GMM	90 %	85 %	87 %
KNN	86 %	88 %	87 %

Les scores sont très supérieurs à ceux obtenus précédemment, car les expériences sont ici effectuées sur le corpus de segments cibles, plutôt homogènes car identifiés par le système initial comme des sons d'eau. Nous pouvons voir que les trois classifieurs obtiennent des scores très similaires. Ce phénomène avait déjà été observé dans [TSGM10]. Les SVM étant légèrement meilleurs que les autres classifieurs, nous les utilisons par la suite.

3.6.3 Résultats

Pour conclure cette étude, nous présentons une évaluation de deux systèmes, qui diffèrent de par l'existence ou non de l'étape de segmentation en segments cibles.

Classifieur sans segmentation

Nous avons effectué une expérience avec le classifieur choisi, le modèle SVM, sur la totalité des 20 vidéos du corpus IMMED. Cette expérience correspond donc à l'utilisation de l'étape de classification de notre système hiérarchique privé de l'étape de segmentation en segments cibles. Cette expérience permet de tester à nouveau sur notre corpus une approche classique basée sur de l'apprentissage automatique.

Dans cette expérience, les signaux sonores sont découpés en sous-segments d'une seconde, puis groupés par vidéo. Ces sous-segments sont annotés come précédemment en fonction de la vérité terrain. Nous utilisons également un ensemble de 18 vidéos en apprentissage et de deux vidéos en test, et nous répétons ce processus 10 fois.

Le tableau 3.5 présente les scores de ce système SVM-Gammatone sur tout le corpus.

Les scores obtenus par le système SVM-Gammatone sur le corpus IMMED sont à nouveau inférieurs à ceux de notre système initial basé uniquement sur la couverture spectrale. Cette expérience confirme une fois encore l'intérêt de notre approche par rapport aux méthodes classiques.

TABLE 3.5 – Classification utilisée de manière autonome.

	Precision	Rappel	F-mesure
SVM-Gammatone	48 %	86 %	**53 %**
Système initial	54 %	83 %	66 %

Ces scores sont assez similaires, mais légèrement supérieurs, à ceux obtenus par une modélisation GMM présentés dans la section 3.4. Une comparaison plus précise est délicate car les protocoles expérimentaux sont assez différents. En effet, le système GMM décrit précédemment effectue une classification par trame suivie d'un post-traitement, alors que le système SVM-Gammatone classe des segments de 1 seconde sans post-traitement. Toutefois, les résultats de cette comparaison vont dans le sens d'études antérieures qui considèrent les gammatones mieux adaptés aux sons environnementaux que les MFCC [VA12]. Il semblerait également que les SVM soient mieux adaptés à cette problématique binaire que les GMM.

Résultats du système hiérarchique

Nous allons maintenant dévoiler les scores de notre système hiérarchique sur les 20 vidéos du corpus IMMED : ce système obtient 82% de F-mesure. Ainsi, une grande majorité des flux d'eau de notre corpus sont détectés, sans que le système produise beaucoup de fausses alarmes. Par rapport à notre précédente étude, les scores sont améliorés de 16%.

Le tableau 3.6 présente les scores de notre système hiérarchique et rappelle les scores précédents sur les 20 vidéos du corpus IMMED.

TABLE 3.6 – Résultats du système hiérarchique.

	Precision	Rappel	F-mesure
Système hiérarchique	79 %	86 %	**82 %**
Segmentation seule (système initial)	54 %	83 %	66 %
Segmentation seule (système BestRecall)	28 %	93 %	44 %
SVM-Gammatone	48 %	86 %	53 %

3.6.4 Conclusion

Le système hiérarchique obtient des scores bien supérieurs aux systèmes précédemment testés. Sa F-mesure dépasse de 15% celle de notre précédent système. Ces résultats sont très encourageants et pourraient permettre une détection d'eau assez robuste au sein de notre projet IMMED. Toutefois, il faut souligner que ce nouveau système intègre de l'apprentissage ce qui présente quelques inconvénients, comme une sensibilité importante aux variations de contexte.

De manière plus générale, les résultats de notre système hiérarchique soulignent l'intérêt d'une segmentation préalable à la classification pour les corpus enregistrés dans des conditions bruités. L'étape de segmentation, basée sur un descripteur robuste, permet ici de réduire considérablement la variété et l'hétérogénéité des données. L'étape de classification, analogue aux systèmes état de l'art utilisés sur des corpus plus maîtrisés, donne par la suite des résultats tout à fait exploitables.

Cette approche semble ainsi généralisable à la détection d'évènements sonores dans des corpus bruités, comme elle l'a déjà été par exemple dans [ZDC06].

Par ailleurs, les scores de rappel de notre système semblent ne pas pouvoir dépasser 94%. Comme nous l'avons expliqué, cette limitation s'explique par le fait que certains sons d'eau de notre corpus annoté par l'audio et la vidéo, ne comportent pas de flux d'eau. L'écoute du corpus IMMED révèle une variété importante de sons d'eau. Par exemple des sons d'éclaboussures lors de l'activité *faire la vaisselle* sont acoustiquement très différents des sons de flux d'eau décrits dans ce chapitre.

Dans ce contexte, l'annotation est délicate. D'une part, il semble impossible d'annoter de manière différente les flux d'eau et les éclaboussures, tellement la limite entre ces deux types de sons est difficilement identifiable : ceux-ci sont d'ailleurs régulièrement superposés.

D'autre part, ces flux d'eau sont parfois à peine audibles. Si l'utilisation de la vidéo a permis de gagner beaucoup de temps lors de l'annotation, elle tend également à nous faire annoter ce que nous voyons plutôt que ce que nous entendons. Par exemple, le flux d'eau, visible à l'écran est parfois masqué pendant de longues secondes par d'autres bruits. Il est ainsi très délicat d'effectuer une annotation temporelle objective. Pour ces deux raisons, l'annotation en son d'eau est dépendante d'une expertise de l'annotateur à pouvoir identifier correctement. Pour évaluer cette difficulté de manière plus objective, nous proposons une expérience perceptive.

3.7 Perception d'extraits du corpus IMMED

3.7.1 But

Le but de cette expérience est de montrer la difficulté d'annoter les sons d'eau sur notre corpus. Elle permet de plus une comparaison entre les performances réalisées par la machine et les décisions humaines sur les extraits choisis, et ainsi à relativiser les scores de la machine dans des problématiques où la reconnaissance parfaite est difficilement envisageable.

3.7.2 Sélection sonore

Nous avons sélectionné manuellement 21 extraits audio de 5 secondes dans les 20 vidéos utilisées précédemment. Ces extraits ont été choisis selon les problématiques énoncées et les résultats du système initial (fausses alarmes, faux rejets, et corrects).

Ces sons peuvent ainsi correspondre à :

- des bruits d'objets acoustiquement proches des flux d'eau (manipulation de sac plastique, passage d'un avion, aspirateur, etc),
- des extraits annotés comme son d'eau où le flux est peu ou pas présent,
- des sons d'eau correctement identifiés par le système.

Les 21 extraits présentant des sonies très différentes, ils ont été normalisés.

3.7.3 Protocole

Une application a été réalisée avec le logiciel MaxMsp [18] pour que chaque participant écoute les différents extraits de manière autonome. La figure 3.18 illustre une des premières fenêtres de cette application.

18. http://cycling74.com/

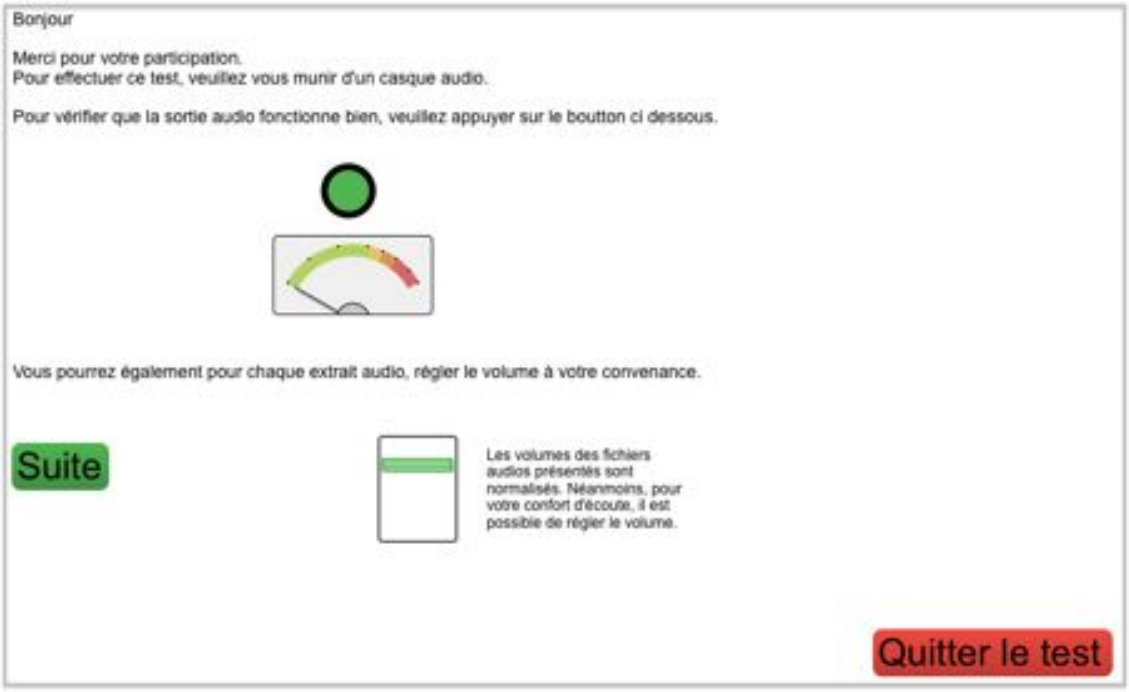


FIGURE 3.18 – Fenêtre de démarrage du test.

La fenêtre 3.19 présente la consigne de ce test. Les sons sont alors présentés de manière aléatoire et différente pour chaque participant. Pour chaque son, le participant doit répondre à la question :

« Entendez-vous de l'eau couler dans cette scène sonore ? ».

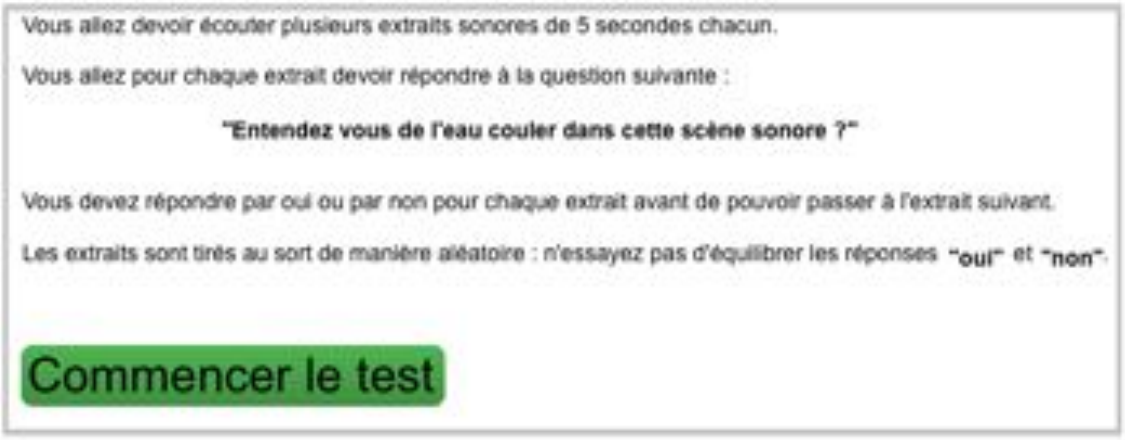


FIGURE 3.19 – Fenêtre de consigne.

Chaque fenêtre d'écoute dispose d'un curseur permettant d'augmenter ou de diminuer l'intensité du son. Il est également possible de réécouter l'extrait autant de fois que nécessaire. La

réponse à la question entraîne la fermeture de la fenêtre et l'ouverture de la fenêtre suivante (voir figure 3.20).

FIGURE 3.20 – Fenêtre d'écoute d'un extrait.

3.7.4 Expérience

Les tests ont été effectués au casque, avec des participants autonomes. 21 participants ont effectué l'expérience. Ces participants travaillent dans notre laboratoire, et trois d'entre eux avaient une connaissance directe de la problématique.

3.7.5 Résultats

Par rapport à l'annotation audiovisuelle, les participants ont effectué en moyenne 79% d'étiquetage correct. Le taux de réponse commune entre participants est de 86%. Sur ces mêmes extraits, notre système initial de détection de flux d'eau (présenté dans la section 3.4) arrive à 52% d'étiquetage correct (selon la majorité du segment annoté).

3.7.6 Discussion

Le protocole et la mise en place de cette expérience peuvent être discutés. Nous pouvons par exemple nous interroger sur l'impact de la normalisation des sons. D'autres expériences plus maîtrisées seront effectuées sur la perception des sons de liquide dans le chapitre 5. La mise en place de cette étape de perception nous a pourtant permis de tirer les conclusions suivantes.

Certains sons annotés comme sons d'eau n'ont pas été reconnus par les participants, probablement car leur sonie était trop faible par rapport aux bruits environnants. Ainsi, les scores de reconnaissance obtenus par les humains permettent de relativiser ceux obtenus par la machine, car la reconnaissance parfaite ne semble pas atteignable. Cette expérience nous a ainsi permis de vérifier que l'annotation audiovisuelle effectuée sur le corpus IMMED n'est pas idéale par rapport à la détection de sons d'eau. Un autre type d'annotation semble toutefois difficile et très

coûteux. Le taux de réponse commune entre participants, inférieur à 100% montre également la subjectivité de l'annotation.

Par ailleurs la mise en place de ce test nous a permis de vérifier une hypothèse. Certains sons de la vie quotidienne sont facilement identifiés par des humains comme des sons d'eau, alors qu'ils ne présentent pas le flux d'eau décrit dans ce chapitre. Ces sons apparaissent pourtant dans les activités de la vie quotidienne, par exemple lorsqu'un patient fait la vaisselle et que le robinet est fermé. Cette considération nous amène à remettre en cause notre modèle basé uniquement sur la reconnaissance de flux d'eau.

3.8 Conclusion

Nous avons dans ce chapitre abordé la problématique de la reconnaissance sonore de flux d'eau. Dans un premier temps, nous avons présenté différentes applications et méthodes de reconnaissance de flux d'eau. Si ces méthodes présentent des démarches très originales et un intérêt scientifique certain, les algorithmes proposés ne sont pas particulièrement adaptés à l'acoustique des sons d'eau. De plus, ces méthodes se fondent sur de l'apprentissage automatique qu'il est difficile d'appliquer sur les données extrêmement variées de notre projet IMMED. Des expériences menées à partir de systèmes état de l'art ont démontré la difficulté de l'utilisation de ces approches.

Nous avons présenté un système développé à partir d'un nouveau descripteur : la couverture spectrale. Ce descripteur a la particularité de présenter une valeur élevée pendant les sons de flux d'eau tout en restant relativement robuste à la présence de parole. Le système proposé obtient des résultats satisfaisants pour la reconnaissance de flux d'eau. Il permet également de détecter l'utilisation de l'aspirateur.

Ce système initial a donc été intégré au corpus IMMED et permet d'inférer sur les activités du patient, notamment à partir du son d'eau. Il a de plus été amélioré par une étape de classification via une approche SVM-Gammatone. Ce système hiérarchique obtient un score de 82 % sur un corpus de plus de 7 heures de vidéo.

Enfin, l'élaboration pratique de ces systèmes de reconnaissance de flux d'eau a ouvert la voie à de nouvelles considérations. Tout d'abord, de nombreux sons du quotidien, reconnus par les humains comme sons de liquide, ne comportent pas de flux d'eau. Ces sons ne peuvent être détectés par notre système, mais sont pourtant présents dans les activités liées à l'eau. Ils correspondent par exemple à des mouvements d'eau, et à des gouttes. Ces différents types de son, facilement identifiables, nous amènent à une interrogation plus précise sur l'origine acoustique des sons d'eau. Le chapitre suivant abordera donc ces sons d'eau d'un point de vue acoustique et vibratoire.

Chapitre 4

Reconnaissance de sons d'eau à partir de modèles physiques

Résumé du chapitre : Selon différentes recherches en acoustique, les sons d'eau sont principalement produits par la vibration de bulles d'air dans l'eau. Ces études théoriques, associées aux analyses de signaux réels, nous ont permis de mettre au point une approche basée sur la reconnaissance de zones temps/fréquence localisées. Des expériences montrent la validité de cette approche pour différents types de sons d'eau. Le système créé est complémentaire avec le système de reconnaissance de flux d'eau décrit dans le chapitre précédent, mais ne peut s'y substituer.

4.1 À l'origine des sons d'eau

4.1.1 Observations préalables

Certaines notions d'acoustique, par exemple les notions liées à la vibration d'une corde, sont relativement familières à un public audiophile. Par contre, s'il semble clair que les mouvements d'eau produisent du son, les phénomènes impliqués dans cette production semblent très peu connus du grand public.

Les expériences précédentes nous permettent de supposer que différents types de sons peuvent être reconnus par des auditeurs comme « son d'eau ». Néanmoins, le point commun entre ces sons n'est pas facilement identifiable. Nous avons ainsi vu dans la partie précédente que le flux d'eau, décrit comme bruité et continu, n'est pas présent dans toutes ces manifestations sonores.

À l'opposé des flux d'eau, certains sons d'eau semblent être composés d'éléments de courte durée, et parfois très localisés en fréquence. Nous pourrions qualifier ces évènements d'« évènements discrets ». La figure 4.1 nous montre ainsi un spectrogramme issu de l'activité *faire la vaisselle* enregistrée dans les conditions du projet IMMED. Le son d'eau a été très facilement reconnu dans cet extrait par les participants à notre expérience.

Sur cet extrait, les fréquences du spectre ne sont pas présentes de manière continue comme c'est le cas pour le flux d'eau. Nous pouvons par contre apercevoir des petites zones temps/fréquence de forte énergie.

Ces zones localisées se retrouvent dans d'autres sons associés par les auditeurs à l'élément liquide : par exemple, un son de robinet faiblement ouvert. Afin de modéliser ces évènements discrets, nous nous sommes tournés vers l'acoustique pour préciser les phénomènes mis en jeu dans la production de son d'eau.

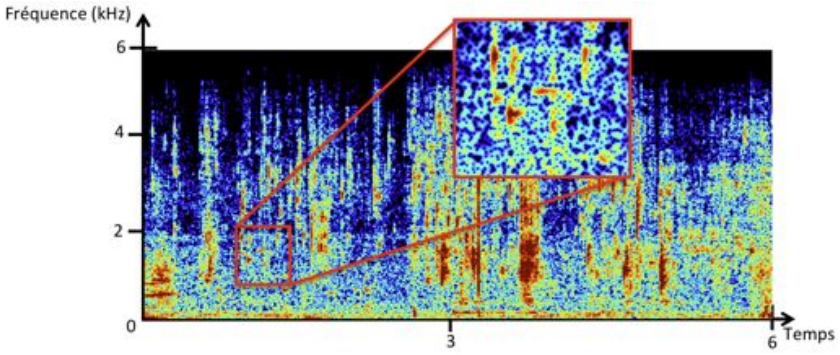


FIGURE 4.1 – Spectrogramme issu de l'activité *faire la vaisselle*.

4.1.2 Origine

D'un point de vue physique, le phénomène à l'origine des sons d'eau est connu depuis longtemps. Il est ainsi admis depuis presque un siècle que l'eau seule ne produit presqu'aucun son [Bra21]. Dans un chapitre traitant des sons d'eau (illustré figure 4.2), Bragg décrit ainsi l'expérience suivante :

FIGURE 4.2 – Extrait du livre *The world of sound* paru en 1921 [Bra21].

In a recent experiment at the Zoological Gardens, observers were stationed round a tank into which they lowered very delicate listening instruments. Fish were thrown in by a keeper, and diving birds went in after them. Not a sound was audible as the birds darted about underneath the water, except when one or two very small air bubbles, carried down by the feathers of one of the birds, came to the surface and burst.

Dans cette expérience effectuée sur un plan d'eau où nagent des poissons et des oiseaux, le seul son capté par des instruments de mesure provient d'une bulle d'air piégée dans l'eau et remontant à la surface.

La plupart des sons impliquant directement l'eau semblent ainsi produits par la vibration de cavités d'air dans l'eau. Plusieurs phénomènes illustrent cette assertion. Par exemple, des

oscillations de faible amplitude à la surface d'un plan d'eau ne sont en général pas audibles. Par contre, si elles sont assez fortes pour créer des vagues qui se cassent et emprisonnent de l'air, elles produisent alors un son tout à fait perceptible.

La photographie de la figure 4.3 illustre ce phénomène assez familier : le son apparait quand une vague se casse.

FIGURE 4.3 – Cassure d'une vague emprisonnant de l'air.

Des propriétés similaires sont observables quand un objet solide tombe dans l'eau. Une chute assez rapide crée une cavité d'air qui se scinde en une multitude de bulles. De même la chute d'une goutte d'eau dans l'eau va former une ou plusieurs bulles d'air (voir figure 4.4).

FIGURE 4.4 – Chute d'une goutte d'eau dans de l'eau.

4.1.3 Historique

En terme de vision, les perturbations résultant de la chute d'une goutte d'eau dans de l'eau ont été étudiées depuis longtemps et produisent des images fascinantes [Wor08]. Sur la figure 4.5, nous pouvons ainsi voir la chute d'une goutte d'eau créant une nouvelle goutte.

Au niveau sonore, la compréhension du phénomène a été moins immédiate. Dans son livre *The world of sound*, Bragg cite Sir Richard Paget, qui, s'il n'a pas publié son travail, semble un des premiers scientifiques à avoir soutenu l'idée que le son de l'eau venait principalement des cavités d'air dans l'eau [Bra21]. Ces cavités, assimilées à des résonateurs, sont pourtant bien trop petites pour donner des fréquences de résonnances audibles.

Dans les années 30, Minnaert émet l'hypothèse que les bulles d'air produisent du son, non pas par l'air résonnant dans des cavités rigides, mais par la vibration des parois de la bulle [Min33]. Cette vibration produit des fréquences tout à fait audibles. Ces fréquences sont transmises à la surface de l'eau, puis dans l'air.

Le modèle proposé par Minnaert est publié sous le titre *On musical air-bubbles and the sound of running water*, reste encore aujourd'hui le point de départ de nombreux travaux sur l'acoustique des sons d'eau. Plus récemment, *The acoustic bubble* reprend un certain nombre des contributions du domaine et propose ainsi plus de 1500 références [Lei97].

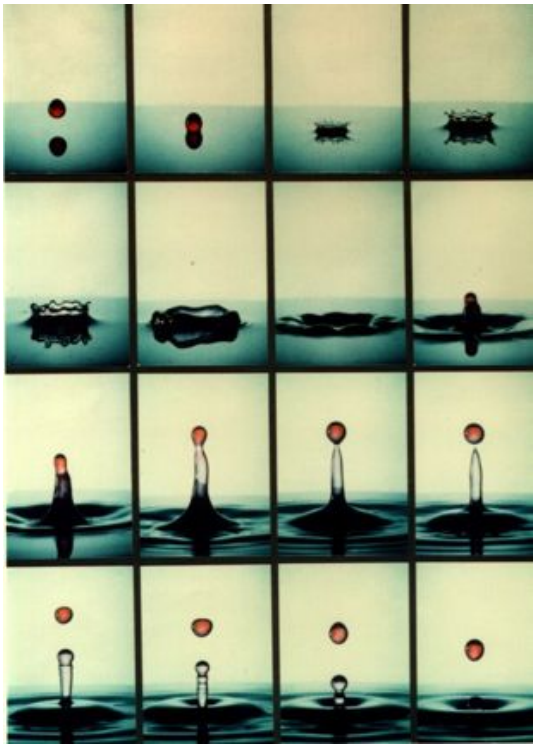


FIGURE 4.5 – Clichés successifs de la chute d'une goutte d'eau colorée dans l'eau.

4.1.4 Synthèse sonore

Ces dernières années, de nombreuses contributions ont été effectuées dans le domaine de la synthèse sonore de son d'eau par modèle physique.

La synthèse par modèle physique vise à créer des sons de manière artificielle à partir des équations de vibrations des matières. Les performances des machines permettent aujourd'hui des temps de calculs tout à fait acceptables pour modéliser et générer ce type de son. Les sons environnementaux, et notamment les sons d'eau, font l'objet d'une attention particulière pour la création de scènes sonores qui peuvent être utilisées dans les jeux vidéo ou le cinéma.

Nous avons ainsi recensé plusieurs travaux utilisant des modèles physiques pour synthétiser les sons d'eau [VdD05, ZJ09, MYH$^+$10]. En particulier, l'article de Kees Van Doel [VdD05] constitue à notre connaissance la première étude en la matière, et a été la source d'inspiration de ce chapitre sur les modèles physiques des sons d'eau.

Ces différents travaux s'appuient exclusivement sur la vibration de bulles d'air pour synthétiser différents types de sons. Ils utilisent notamment les travaux de Minnaert. Le livre de Leighton [Lei97] constitue également une bonne référence, qui propose une modélisation plus complète des phénomènes vibratoires liés aux bulles d'air.

4.1.5 Impacts de gouttes d'eau

Avant de nous focaliser sur le phénomène de vibration des bulles d'air, il faut préciser que la chute de gouttes d'eau fait intervenir d'autres phénomènes physiques.

Impacts de gouttes d'eau dans l'eau

Dans une étude de référence [Fra59], Franz s'intéresse à la chute de corps depuis un gaz vers un élément liquide. Il résume ainsi les différents phénomènes à la base du son, qui apparaissent de manière chronologique dans l'ordre suivant :

1. l'impact du contact entre le corps et la surface liquide,
2. la vibration du corps dans le cas où se dernier est rigide,
3. la pulsation des cavités d'air créées par l'immersion du corps.

L'impact du corps dans le liquide produit en général un son. En effet les déformations de la surface liquide, et éventuellement du corps dans le cas de gouttes se produisent, sur des durées très courtes, à des vitesses supersoniques. Cet impact crée un transitoire d'attaque, d'intensité assez faible.

Nous ne nous attarderons pas sur la vibration du corps car les gouttes d'eau se désagrègent après l'impact. Toutefois, ce type de vibration peut intervenir dans le cas d'un objet solide tombant dans l'eau.

Le dernier phénomène est celui de la vibration de bulles d'air. Il produit de loin le son prédominant. Néanmoins, les expériences de Franz montrent que sous certaines conditions, la chute d'une goutte d'eau dans l'eau peut ne pas provoquer de cavités d'air.

Impacts de gouttes sur des solides

La chute d'une goutte d'eau sur des surfaces solides engendre du son. Les déformations de la goutte lors de l'impact se produisent à des vitesses supersoniques [Rei93]. Acoustiquement, comme dans le cas d'une surface liquide, ces impacts créent des transitoires de faible intensité.

Les interactions avec des surfaces élastiques amènent des contributions significatives des surfaces qui entrent en vibration [ZJ09]. Ce phénomène peut par exemple être utilisé pour déterminer les défauts de construction d'un pont mis en vibration par l'impact de gouttes de pluie [MPG12].

Si nous revenons aux activités de la vie quotidienne, le versement dans un verre en plastique, ou dans un évier en métal va produire un son caractéristique des modes de résonnance du solide. Chaque surface ou objet en interaction va engendrer des sons différents selon sa forme, sa matière, et la manière dont il est excité.

4.1.6 Conclusion

Les travaux d'acoustique nous permettent de trouver des réponses à nos interrogations sur l'origine des sons d'eau. Les sons de liquide viennent principalement de la vibration de cavités d'air. La vibration de ces cavités a été modélisée par les travaux de Minnnaert, qui servent de base à des travaux plus récents d'acoustique. Les études de synthèse de son d'eau s'appuient également sur la vibration de ces bulles.

D'autres interactions interviennent dans la production de son d'eau. Les sons d'impacts présentent une intensité très faible par rapport aux autres sons de vibration. Leur utilisation ne semble donc pas être déterminante pour la reconnaissance de sons d'eau. Ainsi, ces sons ne sont jamais modélisés dans les études de synthèse sonore, qui proposent toutefois la synthèse de nombreux sons différents. Néanmoins, des transitoires d'attaque sont visibles sur les spectrogrammes de certains sons.

Les vibrations de solides en interaction sont difficiles à modéliser. Les possibilités d'interactions sont effectivement très nombreuses, ne serait ce que dans les activités de la vie quotidienne, et dépendent de nombreux facteurs. Nous pouvons de plus supposer que ces interactions solides

apparaissent régulièrement en présence d'interactions gazeuses, et donc que les vibrations de cavités d'air s'ajoutent aux vibrations du solide. La seule vibration de solide ne semble donc pas représentative des sons d'eau. Il semble néanmoins que dans certaines interactions solides, par exemple la montée en fréquence lors du remplissage d'une carafe, elles peuvent être perceptivement fortement associées à l'utilisation de l'eau. Nous évoquerons ce type d'activités dans le chapitre 5.

Au final, la vibration de bulle d'air semble donc être le phénomène physique qui caractérise les sons d'eau. Ce phénomène pourrait ainsi constituer le point commun entre les différents sons de notre test perceptif, et expliquerait pourquoi des sons si différents sont perçus comme « son d'eau ». Nous allons dans la partie suivante nous attacher à décrire le phénomène prédominant et emblématique des sons d'eau : la pulsation des bulles d'air.

4.2 Modèle acoustique de vibration de bulles d'air

Nous allons dans cette section reprendre les premiers travaux de Minnaert sur le son des bulles d'air. Nous ajouterons aux modèles d'origine quelques modifications selon des travaux plus récents d'acoustique décrits dans [Lei97], et selon les approximations utilisées dans les travaux de synthèse sonore.

4.2.1 Système masse-ressort

Dans une première approche, nous pouvons considérer un cas simplifié du problème où :
- la bulle reste sphérique à chaque instant,
- les forces de tension à la surface de la bulle sont négligeables,
- les transformations de l'air dans la bulle sont adiabatiques,
- l'eau est incompressible, et de volume infini.

Nous pouvons alors utiliser une analogie entre la bulle d'air vibrant dans l'eau et un système masse-ressort. L'eau est ainsi représentée par la masse alors que la bulle d'air se comporte comme un ressort (voir figure 4.6). Dans ces conditions, la pulsation propre d'un système masse-ressort de masse m et de raideur k à pour équation : $\omega_0 = \sqrt{\frac{k}{m}}$.

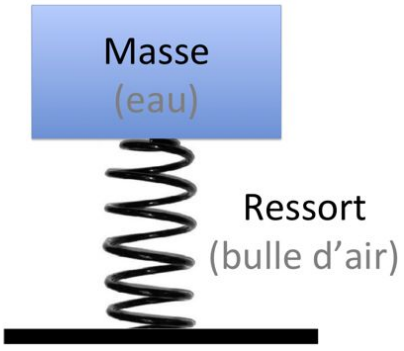


FIGURE 4.6 – Analogie d'une bulle d'air dans l'eau avec un système masse-ressort.

Minnaert démontra dans [Min33] que la fréquence propre de vibration d'une bulle d'air dans l'eau dans les conditions décrites ci-dessus est égale à :

$$\omega_0 = \frac{1}{R_0}\sqrt{\frac{3\gamma p_0}{\rho}} \tag{4.1}$$

où R_0 est le rayon moyen de la bulle, γ est le coefficient de Laplace du gaz parfait, p_0 est la pression moyenne dans la bulle et ρ est la densité de l'eau (ou du liquide). La démonstration de cette équation est disponible en annexe 3.

La bulle d'air peut alors être assimilée à une sphère pulsante de réponse impulsionnelle :

$$p(t) = A_0 sin(\omega_0 t) \tag{4.2}$$

où A_0 est l'amplitude de vibration, déterminée par les conditions initiales d'excitation.

Dans des conditions courantes et en considérant $f_0 = \frac{\omega_0}{2\pi}$, la fréquence de vibration peut être approximée par :

$$f_0 = \frac{3}{R_0} \tag{4.3}$$

Ainsi la fréquence de résonnance d'une bulle de 3 mm sera d'environ 1000 Hz. Par rapport à l'audition humaine, comprise entre 20 Hz et 20 kHz, nous nous intéresserons donc aux bulles dont le rayon est entre 0,15 mm et 15 cm. Par ailleurs, il faut préciser que les bulles de 15 cm, qui produiraient une fréquence de 20 Hz, ne sont pas observables dans la nature. En effet les forces de tension qui s'exercent sur les parois d'une telle bulle d'air dans l'eau sont si fortes que la bulle se scinde immédiatement.

4.2.2 Conditions initiales

Dans le système masse-ressort, un déplacement de la masse crée une condition initiale qui va engendrer l'oscillation du système. Considérons maintenant la formation d'une bulle d'air près de la surface de pression P_0. Les forces de tension de surface qui s'exercent sur les parois de la bulle créent une surpression, égale à $P_o + \frac{2\sigma}{R_0}$ où σ est la tension de surface [ZJ09]. Pour des petites bulles, cette force de tension est très importante. Cette surpression va servir de condition initiale à l'oscillation. La figure 4.7 illustre le changement de pression à l'intérieur de la cavité d'air lorsque la bulle se referme.

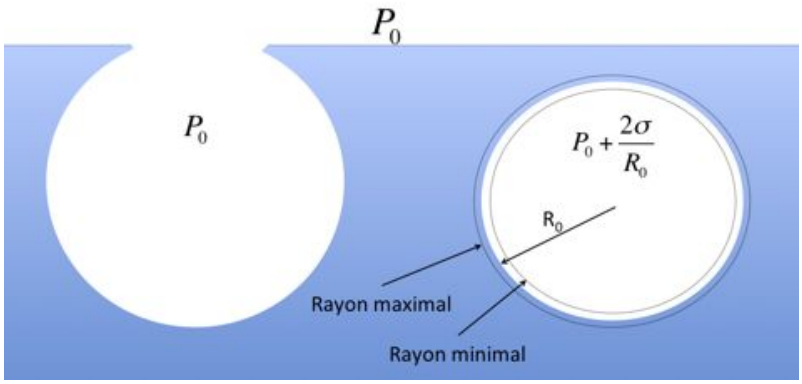


FIGURE 4.7 – Vibration d'une bulle d'air engendrée par les conditions initiales.

4.2.3 Amortissement

Un modélisation plus complète, telle qu'utilisée dans le système masse-ressort consiste à considérer l'amortissement de la vibration.

Dans des conditions réelles, la vibration de la bulle va s'atténuer sous l'effet de forces de dissipation [Lei97]. Ainsi la réponse impulsionnelle de notre bulle d'air peut s'écrire :

$$p(t) = A_0 sin(\omega_0 t) e^{-\delta t} \tag{4.4}$$

L'amortissement δ de la vibration est la somme d'un amortissement de rayonnement qui correspond à l'énergie dissipée par l'onde acoustique, d'un amortissement par conduction thermique et d'un amortissement visqueux. L'amortissement visqueux est négligeable pour des bulles de rayon supérieur à 1 mm. Le détail du calcul de l'amortissement est disponible dans [Lei97]. Nous utiliserons dans la suite l'approximation utilisée dans [VdD05] :

$$\delta = 0.13/R_0 + 0,0072 R_0^{-\frac{3}{2}} \tag{4.5}$$

Il est intéressant de constater que l'amortissement et la fréquence du son augmentent quand le rayon de la bulle diminue. Les sons graves auront donc une durée supérieure à celles des sons aigus. Par ailleurs, ce phénomène semble courant et généralisable à d'autres types de sons. Par exemple dans les sons du quotidien, les sons d'impacts provoqués par des chocs entre deux objets dureront d'autant plus longtemps que la fréquence émise sera grave.

4.2.4 Ascension

Une autre amélioration intéressante de ce modèle consiste à considérer l'ascension d'une bulle d'air dans l'eau. La masse d'eau présente au dessus de la bulle diminue pendant l'ascension de la bulle.

Dans le cas du système masse-ressort où $\omega_0 = \sqrt{\frac{k}{m}}$, nous voyons que la fréquence de vibration du système augmente quand la masse diminue. De la même manière, la fréquence de vibration de la bulle va augmenter à mesure que la bulle se rapproche de la surface.

Ce phénomène est emblématique du son de goutte d'eau tombant dans l'eau. Il se traduit acoustiquement par une augmentation rapide de la fréquence. La figure 4.8 nous montre ainsi le spectrogramme de l'ascension d'une bulle d'air après chute d'une goutte d'eau dans l'eau. La montée en fréquence est visible à partir du temps $t = 0$.

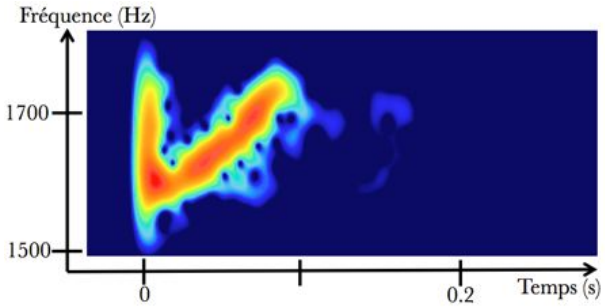


FIGURE 4.8 – Spectrogramme de la chute d'une goutte d'eau dans l'eau.

Par ailleurs, nous pouvons observer sur le spectrogramme de la figure 4.8 un bruit court couvrant une large bande de fréquence au temps $t = 0$. Nous supposons que ce bruit est produit par l'impact de la goutte d'eau dans l'eau.

Concrètement, l'ascension des bulles dépend de plusieurs facteurs dont sa vitesse. La pente de l'élévation en fréquence est assez difficile à mesurer dans des conditions non maîtrisées. Pour modéliser ce phénomène de façon simple, Van den Doel propose l'approximation suivante :

$$f(t) = f_0(1 + \delta\xi t) \tag{4.6}$$

où δ est l'amortissement, et ξ un coefficient ajustable. La fréquence $f(t)$ augmente au cours du temps à partir de sa fréquence de résonnance f_0. À partir de tests d'écoute Van den Doel considère la valeur $\xi = 0,1$ satisfaisante pour produire des sons de bulles réalistes.

D'autre part, une bulle distante d'au moins $10 \times R_0$ de la surface peut être considérée comme totalement immergée. La vibration d'une bulle à cette profondeur ne subira donc pas de changement de fréquence.

4.2.5 Généralisation à des bulles non sphériques

La forme d'une bulle seule dans un environnement stable va converger vers une forme sphérique. Le modèle décrit ci-dessus semble donc constituer une approximation simple et raisonnable. Un modèle plus complet consiste à considérer les bulles non sphériques. Ce type de bulle peut apparaître dans un scénario aquatique plus complexe, faisant intervenir du courant, des vagues ou des turbulences. La figure 4.9 illustre une simulation de synthèse visuelle et sonore pour des bulles non sphériques [MYH+10].

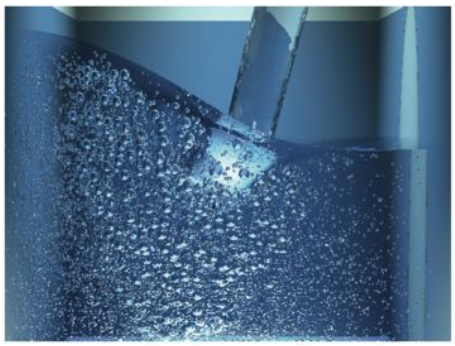


FIGURE 4.9 – Synthèse sonore à partir de bulles non sphériques [MYH+10].

Dans ce cas, la vibration de la bulle produit un son plus complexe composé de fréquences multiples. Selon, [MYH+10], ces fréquences peuvent être approximées pour $n > 1$ par :

$$f_n^2 = \frac{1}{4\pi^2}(n-1)(n+1)(n+2)\frac{\sigma}{\rho R_0^3} \tag{4.7}$$

Les sons provoqués par ces bulles ne sont pas harmoniques.

4.2.6 Dynamique des fluides

Pour des scénario complexes, l'ajout d'un modèle de dynamique des fluides, permet de calculer la position et la forme de chaque bulle pour une action donnée [MYH+10]. Ce type de modèle

permet ainsi de générer les sons correspondants à la chute d'un gros objet ou au déplacement de liquide dans un contenant (voir figure 4.10).

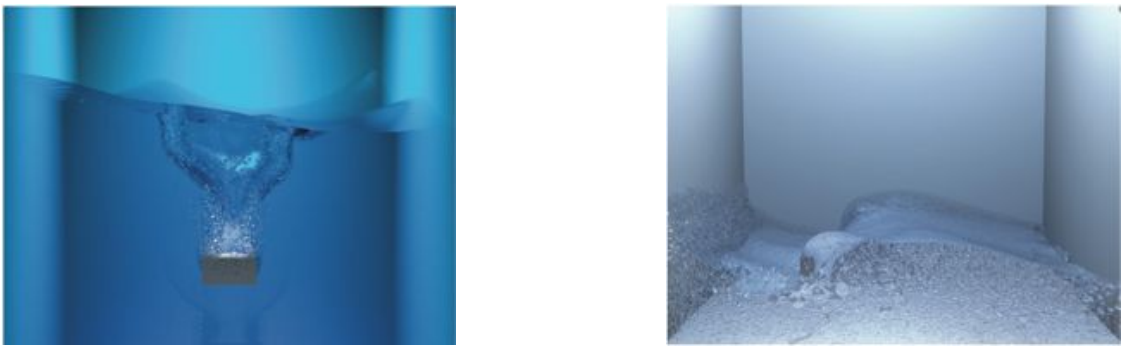


FIGURE 4.10 – Synthèse de sons complexes [MYH+10].

4.2.7 Conclusion

Les études de synthèse montrent qu'un modèle très simple, basé sur un système masse-ressort, permet de générer un son de bulle d'air. Ce modèle permet de synthétiser un son réaliste de goutte d'eau tombant dans l'eau.

À partir du calcul de milliers de bulles d'air de paramètres différents, il est possible de générer une grande variation de sons tels que les sons de pluie, de chute d'eau ou encore de rivière [VdD05].

Une amélioration du modèle est utilisée pour créer des sons de mouvements d'eau, d'éclaboussures, de baignoires et de chute de gros objets [MYH+10]. Par ailleurs, la prise en compte du rayonnement du plan d'eau permet une synthèse réaliste pour un auditeur en dehors de la scène [ZJ09]. Dans tous les cas, ces modèles sont basés sur la vibration de bulles d'air.

Il semble donc envisageable d'utiliser les vibrations de bulles d'air dans l'eau non plus pour synthétiser, mais pour analyser les sons, afin de reconnaître les sons produits par les liquides. Il est toutefois nécessaire d'adapter le modèle théorique issu des analyse acoustiques à nos enregistrements sonores effectués dans la vie réelle.

4.3 Adaptation du modèle à la reconnaissance de scènes sonores

4.3.1 Bulles d'air

Dans des conditions maîtrisées, la vibration émise par la chute d'une goutte d'eau unique dans de l'eau est très visible sur les spectrogrammes (par exemple sur la figure 4.8). Dans de telles conditions, nous pouvons également observer la montée en fréquence provoquée par l'ascension de la bulle.

Comme cette montée rapide en fréquence semble très caractéristique des sons d'eau, il nous a semblé intéressant de détecter ce phénomène. La reconnaissance de son correspond au niveau mathématique à la détection d'un sinus glissant. Nous avons ainsi travaillé sur un modèle mathématique appelée Transformé de Fourier Fractionnaire [GD99, TLW10]. Ce modèle propose une généralisation de la transformée de Fourrier. Suivant un angle donné, cette transformée décompose le signal en somme de sinus glissants. Un balayage de différentes valeurs d'angles permet donc d'observer différentes pentes de sinus glissants.

Toutefois, à notre connaissance, ce modèle a été surtout utilisé dans un cadre théorique, par exemple avec des sons de synthèse. Sur des sons extraits de la vie réelle, il trouve des limitations. Ainsi, il nécessite l'utilisation de fenêtres temporelles de taille très importante, ce qui dans notre cas ne permet pas de détecter les vibrations des bulles.

De plus, comme nous travaillons généralement sur des corpus plutôt bruités, nous nous sommes éloignés de ce modèle théorique. En effet, si nous considérons l'activité *faire la vaisselle* dans notre projet IMMED (figure 4.1), l'observation précise des fréquences résultant de la vibration de bulles d'air n'est pas possible. Par exemple, il n'est pas possible de voir la montée en fréquence produite par l'ascension de la bulle.

Nous pouvons supposer que les conditions d'enregistrements et la complexité des activités réalisées dans l'eau rendent la vibration des bulles difficilement observable. Par contre, nous pouvons également supposer que le phénomène physique de vibration de bulle d'air est à l'origine de zones temps/fréquences de forte énergie visible dans cette figure.

4.3.2 Zones temps/fréquence

Ces zones temps/fréquence de forte énergie pourraient ainsi expliquer pourquoi ce type de son a été immédiatement associé à l'eau dans notre étude perceptive. Cette supposition est étayée par une étude effectuée sur la perception des sons « d'eau courante » [GGWM11]. L'eau courante désigne ici le ruisseau ou la rivière. Maria Geffen et ses co-auteurs s'intéressent ici à la structure spectro-temporelle de ces sons d'eau, qui possèdent des propriétés remarquables.

Une expérience effectuée sur un enregistrement de ruisseau montre que la scène sonore reste perçue comme naturelle même si le son est lu à des vitesses différentes, ces vitesses variant entre 0,5 et 2 fois la vitesse originale. Selon les auteurs, ceci s'explique car la structure du son reste invariante à des transformations spectro-temporelles.

En effet, comme nous l'avons vu l'amortissement de la vibration des bulles d'air dépend de leur fréquence. Le changement de vitesse de lecture contribue donc à changer la durée de ces sons mais également leurs fréquences, ce qui explique qu'ils restent perçus comme naturels.

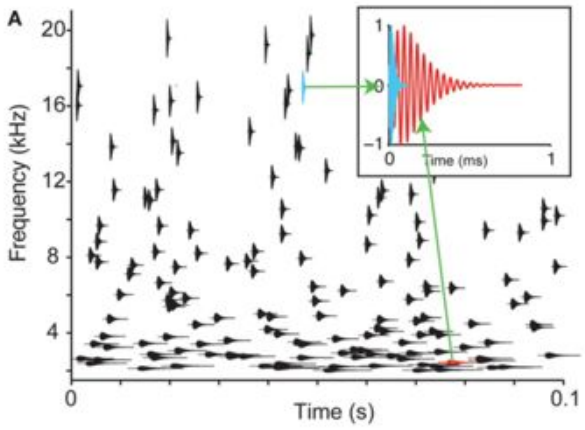


FIGURE 4.11 – Distribution temps/fréquence d'un son élémentaire [GGWM11].

La figure 4.11 illustre une autre expérience effectuée avec des sons de synthèse. Un son de ruisseau est obtenu par superposition de sons élémentaires. La durée de ces sons élémentaires est modifiée selon leur hauteur fréquentielle.

Une bonne paramétrisation de cette relation temps/fréquence permet de rendre ce mélange de synthèse très proche d'un son de ruisseau. Selon les auteurs, un son artificiel serait même perçu comme similaire à l'eau s'il est invariant à la mise à l'échelle. Cette affirmation manque toutefois d'expériences avec d'autres signaux élémentaires pour être considérée comme valable.

Néanmoins, cette étude montre qu'au-delà de la reconnaissance précise des fréquences impliquées dans les vibrations de bulles, la distribution de ces sons élémentaires dans le plan temps/fréquence est d'une importance capitale dans leur perception. Le système auditif possèderait de plus une certaine efficacité à reconnaître des structures invariantes à la mise à l'échelle.

4.3.3 Conclusion

Le modèle physique des sons produits par l'eau est basé sur des ondes sinusoïdales amorties. Toutefois, la complexité des objets intervenant dans les activités liées à l'eau rend la modélisation acoustique de ces activités très délicate.

De façon plus générale, la vibration des bulles d'air engendre des zones de forte énergie très localisées. La taille temporelle de ces zones est liée à leur fréquence. La disposition de ces zones dans le plan temps/fréquence semble être un phénomène perceptivement saillant pour reconnaître ces sons d'eau [GGWM11].

Il nous semble ainsi envisageable de créer un modèle de reconnaissance d'activités liées à l'eau, basée sur la détection de zones de fortes énergies dans le plan temps/fréquence. Ce système pourrait s'appuyer sur les modèles physiques de pulsation de bulles d'air.

Par ailleurs, la chute de goutte d'eau crée un transitoire d'attaque de faible amplitude lors de l'impact. Ces sons d'impact ne semblent pas emblématiques des sons d'eau, ni même nécessaire à la production d'une synthèse réaliste. Toutefois, ils apparaissent dans certains de nos spectrogrammes, et nous utiliserons ainsi l'impact présumé des gouttes à l'origine du son pour repérer précisément le commencement de ces zones.

4.4 Système de reconnaissance

Les travaux d'acoustique décrits précédemment nous permettent d'élaborer un système de reconnaissance de son d'eau basé sur des modèles physiques. La description de ce système et des expériences d'évaluations est publié dans la conférence *International Conference on Acoustics, Speech, and Signal Processing* [GPAO13].

Le schéma de la figure 4.12 présente un aperçu du système. Dans une première phase, nous identifions des candidats potentiels dans un banc de filtre fréquentiel. Puis nous évaluons pour chaque candidat la zone qui l'entoure dans un plan temps/fréquence. Ce traitement nous donne des zones potentielles de vibration de bulle d'air. Afin de détecter les activités liées à l'utilisation de l'eau, nous utilisons une étape finale de post-traitement.

4.4.1 Sélection dans un banc de filtre

Pour reconnaître les sons d'eau nous faisons l'hypothèse qu'à un instant t du son, l'énergie du signal est concentrée dans une petite zone fréquentielle. Cette hypothèse est réaliste avec la vibration d'une ou de quelques bulles d'air successives (voir par exemple les figures 4.1 et 4.11).

Ainsi, selon l'équation 4.1, la vibration d'une bulle unique engendre une seule fréquence à chaque instant. Pour identifier des candidats possibles, nous utilisons un système de banc de filtres pour repérer des zones fréquentielles présentant une énergie importante par rapport au reste du spectre.

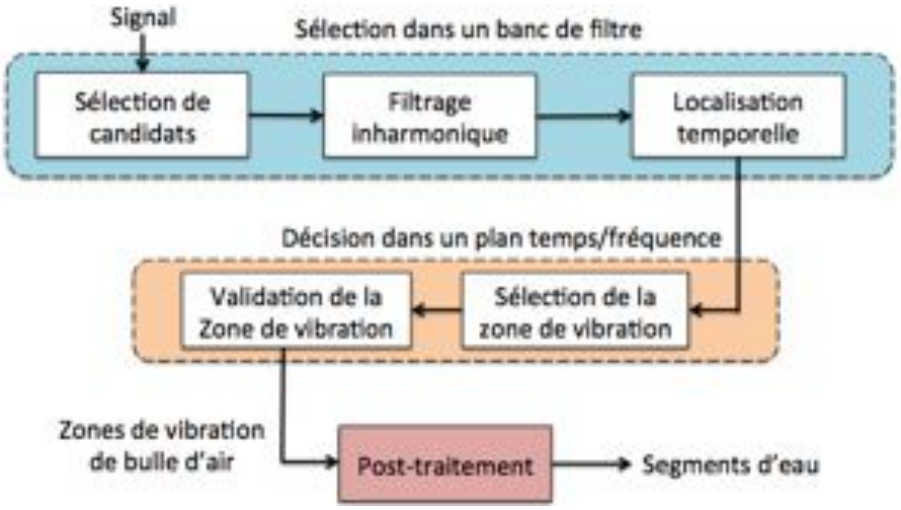


FIGURE 4.12 – Diagramme du système de reconnaissance de son d'eau.

Sélection de candidats

Nous utilisons ainsi des bandes de fréquence de 200 Hz avec un recouvrement de 100 Hz. L'énergie E_{bin} de chaque point fréquentiel est ensuite normalisée par rapport à l'énergie E_{trame} de la trame. Nous sélectionnons donc des points de forte énergie comme candidats grâce à un seuil $T_{candidat}$ selon la formule :

$$\frac{E_{bin}}{E_{trame}} > T_{candidat} \tag{4.8}$$

Par ailleurs, les bulles d'air sont limitées en taille car les forces de tensions poussent les grosses bulles au fractionnement. L'équation 4.3 nous donne une relation entre la taille de la bulle et la fréquence : $f_0 = \frac{3}{R_0}$.

Si certains sons graves produits par des grosses cavités d'air peuvent parfois apparaître dans la nature, nous allons nous concentrer sur les bulles de diamètre plus petit. En effet, si de grosses cavités d'air peuvent être engendrées par le jet d'une grosse pierre dans une mare, il semble raisonnable de penser que cette activité n'apparait que rarement au sein du corpus IMMED. De plus ce type d'activité produit obligatoirement d'autres bulles de diamètre plus petit.

Nous allons nous focaliser sur les bulles dont la fréquence émise est supérieure à f_{min}. Nous supprimons ainsi les points candidats de fréquence inférieure à $f_{min} = 800$ Hz.

Filtrage harmonique

Dans le cas du modèle simple de la sphère pulsante comme dans les modèles plus compliqués, la vibration de bulle d'air ne produit pas de partiels harmoniques (voir les équations 4.1 et 4.7). Dans cette étape de filtrage, nous supprimons les candidats détectés sur les trames harmoniques du signal. Cette étape nous permet d'être robuste aux formants de la parole, et éventuellement à d'autres sons harmoniques.

Nous supposons que la parole est caractérisée par une fréquence fondamentale assez basse, qui est en dessous de f_{min}. Nous utilisons l'algorithme *Yin* [DCK02] pour déterminer la fréquence fondamentale f_0 des trames du signal.

Par ailleurs, le *Yin* produit des erreurs dans des contextes bruités. Ainsi, dans le cas de segments bruités le *Yin* n'est pas assez fiable pour que nous puissions valider la fréquence fondamentale détectée. Nous utilisons une mesure de périodicité $p(t)$ fournie par l'algorithme pour vérifier la fiabilité de la fréquence fondamentale trouvée. Nous fixons un seuil de périodicité du segment p_{min}, pour supprimer les candidats tels que :

$$f_0(t) < f_{min} \quad et \quad p(t) > p_{min} \tag{4.9}$$

Localisation temporelle

Afin de déterminer une zone temps/fréquence autour du candidat, nous localisons le début de la zone sur la même plage de fréquence. Le début de zone est affecté au point d'énergie minimale entre le candidat et un point situé 100 millisecondes avant le candidat. Nous supposons alors que la vibration de la bulle d'air commence à cet instant. Cette étape nous permet d'ajuster temporellement notre zone temps/fréquence.

Par ailleurs, si le minimum trouvé est le candidat, nous supprimons le candidat. En effet, dans ce cas, l'énergie est constante, ce qui ne correspond pas à l'amortissement de notre modèle de vibration de bulle d'air.

4.4.2 Décision

Après avoir sélectionné les candidats et repéré temporellement le début de la zone dans une bande de fréquence, nous allons considérer les candidats dans leur globalité spectrale.

Sélection de la zone de vibration

Temporelle : Pour calculer la taille temporelle de notre zone, nous utilisons l'équation du système masse-ressort amorti (4.4 : $p(t) = A_0 sin(\omega_0 t) e^{-\delta t}$). La vibration diminuant de manière exponentielle, nous fixons un seuil ϵ à partir duquel la vibration est considérée négligeable. Nous recherchons un temps t tel que :

$$|A_0 sin(2\pi f t) e^{-\delta t}| < \epsilon \tag{4.10}$$

Cette dernière équation est vrai si $t > \frac{ln(\epsilon/A_0)}{\delta}$.

Pour calculer l'amortissement δ de la bulle, nous utilisons l'approximation utilisée dans [VdD05] décrite dans la partie 4.2.3 :

$$\delta = 0.13/R_0 + 0,0072 R_0^{-\frac{3}{2}} \tag{(4.5)}$$

L'amortissement δ dépend ici du rayon de la bulle. Comme le rayon de la bulle peut être facilement calculé à partir de sa fréquence ($R_0 = \frac{3}{f_0}$), nous pouvons calculer un amortissement δ pour chaque fréquence f_0.

Au final, en fixant un seuil ϵ exprimé en décibels, et en considérant $A_0 = 1$, nous pouvons déterminer un temps t de vibration dépendant de la fréquence de chaque candidat.

Fréquentielle : D'après l'état de l'art [Fra59], l'augmentation en fréquence de la vibration est difficilement prédictible. Nous utilisons donc une zone de fréquence fixe assez large pour être robuste aux changements de fréquence. Nous fixons une largeur fréquentielle de 500 Hz autour de chaque candidat.

Validation de la zone de vibration

À l'issue de l'étape de sélection, nous repérons dans le spectrogramme une zone rectangulaire autour de chaque candidat. Afin de valider cette zone de vibration, nous devons nous assurer que la potentielle vibration d'une bulle d'air se limite temporellement à la zone détectée. Ce traitement nous permet ainsi d'être robuste aux sons continus qui pourraient être détectés.

Nous définissons alors deux autres zones : la zone *pre-vibration* et la zone *post-vibration*. En considérant la durée de vibration des bulles d'air, nous supposons que ces deux zones présentent beaucoup moins d'énergie que notre zone de vibration. La figure 4.13 nous montre la sélection de ces zones.

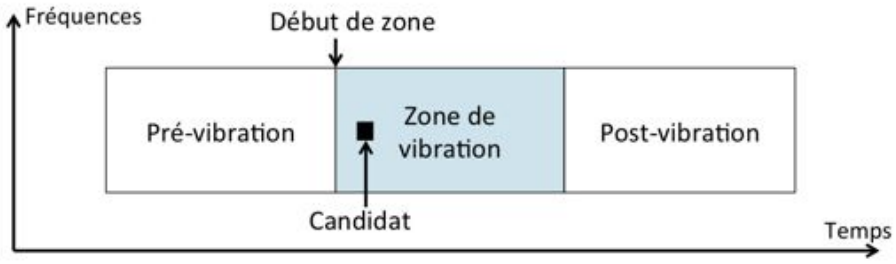


FIGURE 4.13 – Sélection de zones autour du candidat.

Nous estimons par ailleurs que la localisation temporelle du début de zone est plus précise que la longueur de la zone de vibration, qui dépend concrètement de nombreux facteurs. Nous utilisons donc une condition relachée sur la zone *post-vibration* par rapport à celle de la zone *pre-vibration*.

Au final, pour $E_{vibration}$ l'énergie de la zone de vibration , nous validons la zone si l'énergie de la zone de *pre-vibration* E_{pre} et l'énergie de la zone de *post-vibration* E_{post} sont telles que :

$$E_{pre} < 0,5 * E_{vibration} \quad et \quad E_{post} < 0,8 * E_{vibration} \tag{4.11}$$

Enfin, pour éviter d'obtenir un recouvrement entre différentes zones validées, nous supprimons tout candidat trouvé dans une zone de vibration validée.

4.5 Développement

4.5.1 Constitution du corpus

Un corpus de développement pour notre système a été constitué. Il est composé de deux segments de 5 secondes extraits du corpus IMMED. Le premier est un extrait de l'activité *faire la vaisselle* contenant de nombreux sons de gouttes et de mouvements d'eau. Le deuxième est un extrait contenant principalement de la parole.

Nous avons ajouté à ces deux extraits 20 sons extraits du projet *Freesound* [AFF+11]. Parmi ces sons, 15 font intervenir des sons d'eau : *verser, dans un verre, robinet ouvert, saut dans*

une piscine, rivière, eau bouillante. Les autres sons ne font pas intervenir de sons d'eau, mais comportent des évènements sonores qui peuvent éventuellement produire des fausses alarmes : *réveil, sonnette, ouverture et fermeture de porte*.

Tous ces extraits sonores ont été convertis dans un format Wave quantifié à 16 bits et échantillonné à 16 kHz.

4.5.2 Seuillage

Les spectrogrammes de notre système sont calculés par l'algorithme FFT, après un fenêtrage de *Hamming* sur 512 points. Nous avons fixé nos seuils sur ce corpus de développement et obtenons :

- $T_{candidat} = 0,15$
- $p_{min} = 0,6$

4.5.3 Résultats qualitatifs

Nous pouvons voir sur la figure 4.14 le résultat de notre système sur certains fichiers du corpus de développement. Les rectangles noirs correspondent aux zones de vibration validées par notre système. Les figures de gauche correspondent à des sons faisant intervenir l'eau, celle de droite à des sons pouvant créer de fausses alarmes.

La figure 4.14-a illustre les résultats sur 3 secondes d'extrait de l'activité *faire la vaisselle* du corpus IMMED. Nous pouvons voir que plusieurs zones sont détectées sur cet extrait où interviennent de nombreux sons d'eau. La 4.14-b présente également un extrait du corpus IMMED, sans bruits liés à l'eau. Dans cet extrait, notre système ne produit pas de fausses alarmes, malgré une forte présence de voix.

La figure 4.14-c illustre une goutte d'eau tombant du robinet. Il est intéressant de constater que la vibration de bulles d'air est difficilement observable en regardant le spectrogramme. Nous apercevons par contre très facilement deux impacts successifs. Deux zones successives sont détectées par le système, la deuxième zone étant plus aiguë que la première. Nous pouvons supposer que la chute d'une goutte d'eau crée dans ce cas une nouvelle goutte (comme cela est illustré sur la figure 4.5 page 76). Chacune de ces gouttes crée une bulle d'air : la deuxième goutte, plus petite, créant une bulle de diamètre inférieur. Notre système semble donc capable de capter des détails difficilement visibles sur le spectrogramme.

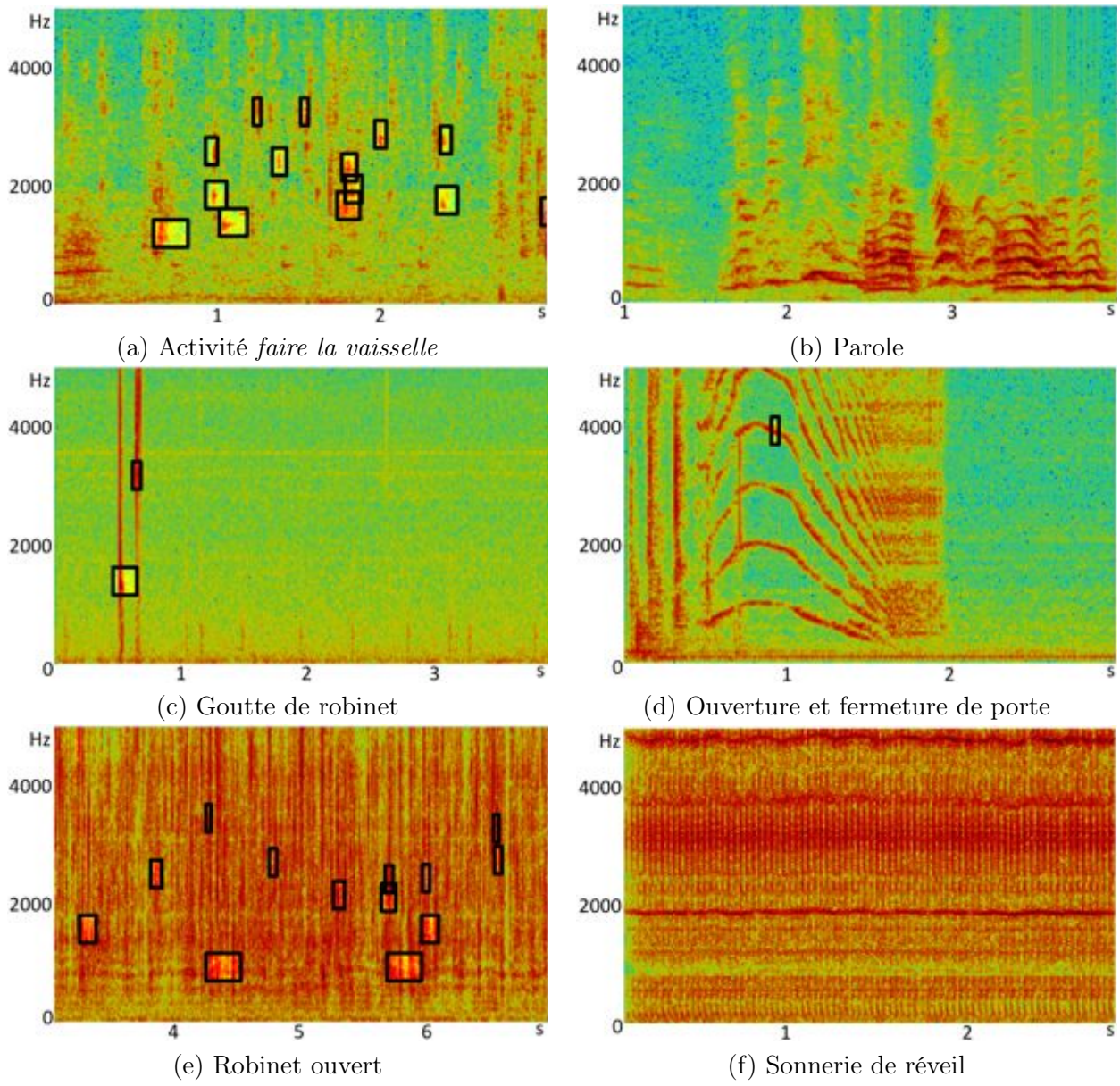
La figure 4.14-d nous montre l'ouverture et la fermeture d'une porte qui grince. Ce son harmonique présente une fréquence fondamentale élevée. De plus, il varie en fréquence. Ici, une zone temps/fréquence de forte énergie a été détectée aux environs de 4000 Hz. Une étape de post-traitement peut supprimer ce type de fausses alarmes isolées.

La figure 4.14-e illustre un son de robinet ouvert. Ce son est très bruité, et semble proche du flux décrit dans le chapitre précédent. Pourtant, même dans ces conditions, il semble que certaines zones de vibration se détachent et sont repérées par notre système.

Enfin, la figure 4.14-f illustre un son de réveil : aucune fausse alarme n'a été repérée dans cet extrait.

4.5.4 Post-traitement

À l'issue de ce développement, une étape de post-traitement a été ajoutée à notre système. Le post-traitement consiste à supprimer les zones isolées dans le signal. Nous supposons en effet que les zones isolées correspondent à des fausses alarmes, comme sur la figure 4.14-d. En effet les activités liées à l'eau semblent en général produire un nombre de bulles d'air important. Toutefois



(a) Activité *faire la vaisselle* (b) Parole

(c) Goutte de robinet (d) Ouverture et fermeture de porte

(e) Robinet ouvert (f) Sonnerie de réveil

FIGURE 4.14 – Reconaissance de zone de vibration sur un corpus de développement.

cette étape de post-traitement peut empêcher notre système de détecter les chutes de gouttes d'eau isolées, illustrées par exemple sur la figure 4.14-c.

Nous considérons une fenêtre glissante de durée s_t. Nous supprimons les zones dont le nombre est toujours inférieur à s_n. Nous espérons ainsi obtenir les sons d'eau, ou les activités liées à l'eau à l'issue de ce post-traitement.

Les seuils suivants on été fixés sur le corpus de développement :

- $s_t = 2,5s$
- $s_n = 5$

4.6 Expériences

Afin de valider notre système, nous avons proposé deux expériences. Par souci de reproductibilité nous avons effectué notre première expérience sur un corpus accessible par d'autres chercheurs en utilisant les enregistrements de *BBC-Sound_ Effet-Library*. Nous utilisons dans une deuxième expérience le corpus IMMED, qui, pour des raisons de droit à l'image et de secret médical, est consultable uniquement par les membres du projet. Cette deuxième expérience permet de comparer les résultats de ce système avec les stratégies précédentes de détection de son d'eau.

4.6.1 Classification sur un corpus de sons du quotidien

Nous avons utilisé le disque de sons environnementaux *BBC-Sound_ Effet-Library{#3: Household}* [19] pour créer un corpus de test. Ce CD contient 47 fichiers, pour une durée totale d'une heure. Il est composé d'enregistrements d'activités à la maison. Il ne contient pas de parole, ni de bruits extérieurs aux activités. Parmi les activités, 9 enregistrements correspondent à des activités utilisant l'eau, par exemple : *se laver les mains, bain, chasse d'eau, robinet qui goutte, machine à laver, etc.* Les 38 autres correspondent à d'autres activités, parmi elles : *utilisation d'une scie électrique, porte grinçante, alarme, sonneries de téléphone, sonnette.*

L'objectif de cette expérience est de classer chacun de ces enregistrements selon deux classes : les activités liées à l'eau et les autres activités. Nous proposons une évaluation en terme de nombre d'activités reconnues. Le tableau 4.1 illustre les résultats de notre système :

Table 4.1 – Résultat de la classification.

	Son d'eau	Autres sons
Vérité terrain	9	38
Notre système	9	9

Notre système a reconnu tous les enregistrements de son d'eau et a produit 9 fausses alarmes. La F-mesure en terme d'évènements détectés est de 66%. L'une des fausses alarmes est discutable car le fichier correspond à l'activité *faire frire un oeuf* qui présente des éléments liquides.

Parmi les autres fausses alarmes, nous trouvons également les sons d'utilisation d'une scie électrique ou d'une allumette. Il semble ainsi que les fausses alarmes correspondent à des sons composés d'impacts multiples produisant des fréquences aiguës. L'utilisation de ces objets produit des zones temps/fréquence de vibration potentiellement reconnues par notre système.

4.6.2 Détection sur un extrait de la vie réelle

La seconde expérience a été effectuée sur un enregistrement de 21 minutes issu du corpus IMMED. Ce flux audio contient différentes activités d'une personne âgée à la maison. Comme il a été enregistré dans la vie réelle, il contient également d'autres sons, tels que le chant des oiseaux.

Ce fichier contient 82 secondes de l'activité *faire la vaisselle*. Néanmoins, le flux d'eau n'est presque pas audible dans cette activité. Ainsi nos précédents systèmes basés sur la reconnaissance de flux d'eau ne pouvaient pas le détecter. Des extraits de cette activité avaient pourtant été

19. The BBC sound effects library. http://www.sound-ideas.com

facilement reconnus par les participants à notre test comme faisant intervenir l'eau. L'activité engendre en effet différents sons d'eau, dont des mouvements et des gouttes.

Les résultats de notre système, en terme de durée d'évènements sonores correctement détectée, sont visibles dans le tableau 4.2. Une grande partie de l'utilisation de l'eau a été reconnue. Les fausses alarmes sont principalement causées par la manipulation de sacs en plastique, par des chocs sur la caméra, et parfois par des oiseaux.

TABLE 4.2 – Détection sur un extrait de la vie réelle.

	Son d'eau	Autres sons
Vérité terrain	82 s	1171 s
Notre système	63 s	35 s

En terme de durée la F-mesure de cette expérience est de 70 %. Pour comparaison, nous avons effectué un test sur ce fichier avec deux des systèmes proposés dans le chapitre 3. Le système classique GMM/MFCC obtient une F-mesure de 36% sur cet extrait et notre système de détection de flux d'eau approche les 45%. Notre nouveau détecteur a donc permis une amélioration de la reconnaissance d'eau sur ce fichier.

4.7 Conclusion

4.7.1 Vibration de bulles d'air

Nous avons présenté dans ce chapitre un système de reconnaissance de son d'eau basé sur des modèles physiques. Une approche transversale de l'état de l'art sur les sons d'eau nous a permis de préciser les caractéristiques significatives de ces sons. Les sons d'eau semblent principalement créés par la pulsation de bulles d'air dans l'eau. De plus, la disposition des zones temps/fréquence produites par les pulsations semble être un indice perceptif important pour leur reconnaissance.

A partir de ces observations, nous avons élaboré un système de reconnaissance automatique, qui permet de détecter des zones temps/fréquence très localisées dans le spectrogramme. Ces zones peuvent correspondre aux vibrations de bulles d'air dans l'eau. Les équations des modèles acoustiques ont permis d'affiner ce modèle, notamment par le calcul précis de la durée des zones.

4.7.2 Modélisation de l'évènement sonore

Les expériences proposées ont contribué à justifier cette approche. En effet, sur une base de données composée de sons environnementaux, tous les sons de liquide ont été reconnus. Ainsi, un des gros avantages de cette approche, est qu'elle permet de détecter des sons faisant intervenir l'eau issue d'activités très différentes, en se focalisant sur le phénomène physique commun à tous ces sons. Une autre expérience menée sur un extrait du corpus IMMED montre que ce système détecte l'activité *faire la vaisselle* presque entièrement là où d'autres approches ne donnaient pas de résultats satisfaisants.

Contrairement à l'approche sur les flux d'eau développée dans le chapitre précédent, notre système se base sur l'atome génératif de ces sons : la vibration des bulles d'air. En s'attachant à détecter ce phénomène, notre système est potentiellement capable de reconnaître des sons aussi divers que ceux liés à la piscine, au versement dans un verre, à l'utilisation d'un arrosoir, au pressage d'une éponge, ou à l'utilisation d'une serpillière.

Ainsi, cette approche semble adaptée à notre problématique du projet IMMED dans laquelle nous proposons de reconnaître des sons d'eau différents dans des conditions variées pour inférer sur l'activité des patients. Ce système pourrait être utilisé pour d'autres applications où la détection de sons d'eau différents des flux est intéressante, par exemple la détection de chutes dans des piscines.

4.7.3 Limites et complexité

Lorsque nous sommes en présence d'un flux d'eau tel que décrit au chapitre 3, la reconnaissance de son d'eau est moins évidente. Certains exemples sont encourageants. Par exemple, sur la figure 4.14-e, nous pouvons voir que même en présence d'un flot plutôt bruité, la vibration de certaines bulles se détachent acoustiquement du fond sonore. Certaines vibrations de bulles peuvent également ressortir en début et en fin de l'évènement sonore. Malgré ces observations positives, il semble difficile par notre approche de pouvoir détecter certains sons d'eau très complexes, qui ne font pas ressortir de zones temps fréquences particulières. Ces sons correspondent typiquement à des sons de flux d'eau.

Ainsi, au final, nous disposons de deux systèmes de détection qui semblent complémentaires. D'une part le système de détection de flux ne peut reconnaître certaines activités. D'autre part, le système de reconnaissance de zone de vibration est limité par la complexité du signal.

Il est de plus intéressant de constater que les sons détectés par ces deux systèmes sont très différents, voire complètement opposés. Les flux d'eau couvrent une large bande de fréquence et s'étalent dans la durée. Les zones de vibration que nous détectons dans ce chapitre sont courtes et localisées en fréquence. Pourtant, en terme d'acoustique ces types de sons très différents semblent être produits par un phénomène unique : la vibration de bulles d'air. Il semble que le nombre de bulles en vibration soit à l'origine des différences entre ces deux types de signaux. Par ailleurs, l'attaque produite par la chute d'une goutte ou les vibrations solides peuvent également contribuer à la production d'un signal complexe et bruité.

Si le nombre de bulles est le seul facteur entre ces deux catégories de son, nous pouvons nous interroger sur la frontière et le nombre limite de bulles entre ces deux phénomènes. Pouvons-nous distinguer précisément des groupes de sons appartenant à l'une ou à l'autre des catégories ?

Des participants humains ont reconnu facilement des sons de ces deux catégories comme des sons d'eau. Pour répondre à la question précédente, il serait intéressant de savoir si la reconnaissance de ces sons vient de deux processus cognitifs différents qui permettent d'associer chaque catégorie à la même cause : les sons d'eau. La réponse à cette question permettrait éventuellement de valider l'utilisation de deux systèmes pour reconnaître deux catégories de sons produits physiquement par une cause commune. Une étude perceptive est présentée dans le chapitre suivant.

Chapitre 5

Perception des sons de liquide

Résumé du chapitre : La catégorisation est un processus cognitif permettant de regrouper des objets selon des propriétés similaires. Les expériences de classification de sons permettent d'identifier les catégories qui émergent lorsque les personnes se représentent mentalement le monde sonore. Si plusieurs études ont été effectuées sur les sons environnementaux et notamment sur les sons de solides, aucune étude ne mentionne d'analyse spécifique des catégories des sons de liquides. Nous présentons ici des expériences dont l'objectif est d'établir des catégories validées perceptivement. Les résultats d'analyse nous amènent à l'identification de catégories de sons de liquide.

5.1 État de l'art

5.1.1 Introduction

La reconnaissance automatique de sons d'eau dans le cadre du quotidien nous a conduits à utiliser deux approches différentes. La première permet de détecter les sons de flux d'eau (chapitre 3), alors que la seconde approche se focalise sur les sons produits par les vibrations de bulles d'air dans l'eau (chapitre 4).

Ces deux approches sont complémentaires et permettent la détection des différents sons d'eau présents dans les enregistrements du quotidien. Les zones temps/fréquence produites par les bulles d'air, dans le cas par exemple de gouttes tombant dans l'eau, ne peuvent être détectées par une approche basée sur des descripteurs fréquentiels pour détecter les flux d'eau. De même, les flux d'eau, produits par un nombre important de gouttes, sont trop complexes pour être reconnus par notre approche de détection de zones temps fréquence, qui est limitée à un petit nombre de bulles simultanées.

L'utilisation de ces deux méthodes soulève des questions relatives à l'ensemble des sons d'eau entendus dans le quotidien. Est-il nécessaire d'une part d'utiliser deux approches différentes pour détecter les différents types de sons d'eau dans cet ensemble ? D'autre part, ces deux approches sont elles suffisantes pour détecter l'ensemble de ces sons, ou existe t-il des sons d'eau du quotidien que ces approches ne peuvent détecter ?

Une étude précise des différents types de sons d'eau que nous pouvons entendre au quotidien, peut nous permettre de répondre à ces interrogations. Il serait ainsi intéressant de consulter une taxonomie des sons d'eau pouvant être entendus dans son domicile.

Nous avons vu dans le chapitre précédent que ces différents types de sons semblent pouvoir être produits par la même source physique. Ainsi, il est difficile d'établir une dissociation claire entre plusieurs types de sons d'eau par l'étude du modèle physique à l'origine de ces sons.

Par contre, la perception humaine dissocie de manière assez précise deux types de sons résultant de la même source physique. Ainsi, les sons d'impacts ne sont plus discernés isolément dès que leur nombre dépasse une fréquence d'environ 20 Hz. Un marteau tapant sur une surface produira un son considéré comme continu au delà de 20 impacts par secondes, et la fréquence de ce son, d'abord très grave, augmentera avec le nombre d'impacts [PPK01].

Nous pouvons effectuer une analogie entre sons d'impacts/sons continus et vibrations de bulles d'air/flux d'eau. Ainsi, pour identifier les différents types de sons d'eau et justifier les propriétés nécessaires et suffisantes de nos deux approches, nous nous sommes appuyés sur une étude perceptive.

5.1.2 Perception des sons d'eau

La perception sonore est un domaine de recherche qui peut faire intervenir les sciences cognitives, la psychologie expérimentale, les neurosciences, l'acoustique, la linguistique ou encore la médecine. Des études perceptives portant précisément sur les sons d'eau ont déjà été effectuées ces dernières années.

Nous avons cité dans le chapitre précédent (partie 4.3.2) l'article de Maria Geffen et de ses collaborateurs [GGWM11]. Cette étude révèle l'importance de la répartition dans le plan temps/fréquence de zones d'énergie pour la perception des sons d'eau. Les auteurs montrent que la structure du son d'eau est invariante à des transformations spectro/temporelle. Des sons respectant cette structure semblent ainsi identifiés comme sons de liquides.

Une autre étude intéressante a été effectuée sur les sons de remplissage de contenant [CP00]. D'un point de vue acoustique, la fréquence de vibration du contenant varie linéairement avec la longueur vibrante. Le remplissage du contenant fait diminuer cette longueur vibrante et produit une montée en fréquence du son de vibration. Nous pouvons ainsi observer sur la figure 5.1 l'augmentation de certaines fréquences pendant le remplissage.

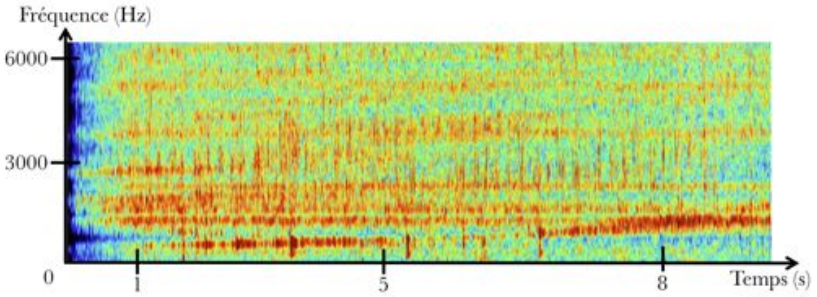


FIGURE 5.1 – Spectrogramme issu du remplissage d'une bouteille en verre.

Le son de cette vibration, qui est ajouté aux « gargouillis » de l'eau, permet d'obtenir des indices sur le volume rempli, comme cela a été démontré par Patrick Cabe et John Pittenger [CP00]. Les expériences de cette étude montrent que ces sons aident sensiblement à évaluer le remplissage d'un contenant. Les participants sont pourtant moins performants lorsqu'ils utilisent le son seul, que lorsqu'il est ajouté aux modalités visuelles et tactiles.

Nous avons vu dans les chapitres précédents que les sons d'eau, même dans le cadre du quotidien, peuvent produire une certaine variété de sons. De plus, notre expérience perceptive décrite dans le chapitre 3, suggère que ces sons variés sont associés facilement à l'élément liquide

par les auditeurs. Malgré la diversité des sons d'eau et leur capacité à être facilement identifiés, il n'existe pas à notre connaissance d'étude perceptive sur les différents types de sons d'eau. Ainsi, devant l'absence de taxonomie de sons d'eau, nous avons ouvert notre recherche à l'ensemble des sons environnementaux.

5.1.3 Perception des sons environnementaux

Différents types d'écoute

Avant de parcourir les différentes études menées sur la perception des sons environnementaux, il faut préciser que notre référentiel a changé. Ainsi les études qui vont suivre se placent généralement du point de vue de l'auditeur. La définition du chapitre 2, où nous considérions les sons environnementaux comme « tous sons autre que la musique ou que la parole » n'est plus suffisante dans ce contexte, car elle considère uniquement la production du son, et non pas la manière dont il est perçu.

Plusieurs études sur la perception des sons environnementaux ont comme point de départ les recherches menées par Gaver [Gav93], qui définit différents types d'écoute. Il décrit ainsi l'écoute musicale (*musical listenning*) et l'écoute de tous les jours (*everyday listenning*) [Gav93]. Dans l'écoute musicale, l'auditeur s'attache aux propriétés perçues des sons comme la hauteur ou le timbre. Ces propriétés sont liées à des paramètres acoustiques tels la fréquence fondamentale ou le centroïde spectral. Dans l'écoute de tous les jours, l'auditeur s'attache à déterminer la cause du son, ce qui lui permet d'inférer sur son environnement.

Selon cette approche, la musique est en général écoutée dans le cadre de l'écoute musicale, et les autres sons selon une écoute de tous les jours. Toutefois, le type d'écoute est ici défini du point de vue de l'auditeur, et non de la production sonore. Un auditeur peut en effet pratiquer l'écoute de tous les jours lorsqu'il entend de la musique, par exemple en identifiant l'instrument jouant une certaine partie musicale lors d'un concert. D'autre part, un son considéré comme environnemental selon d'autres approches, peut être écouté comme de la musique, comme dans le cas de la *musique concrète* [SC67].

Au delà de la musique concrète, les sons environnementaux sont le plus souvent écoutés, ou entendus, lors de l'écoute de tous les jours. Ils nous permettent d'interagir avec notre environnement, de savoir qu'une voiture approche ou qu'un objet est tombé derrière nous. L'écoute de ces sons peut nous permettre d'une part d'inférer sur la source sonore (par exemple la voiture) et ses propriétés (taille, vitesse, modèle, etc). D'autre part, elle peut permettre d'identifier l'action effectuée ou le mouvement de l'objet, par exemple la chute. L'audition d'un son peut ainsi apporter des informations sur l'objet impliqué ou l'action effectuée [Dub00].

Perception causale

Différentes études ont été menées dans le cadre de la perception de la cause des sons afin de préciser le rapport entre le son perçu et la source sonore évoquée. Ces études visent notamment à quantifier les capacités d'identification des propriétés de la source ou des actions effectuées.

Des recherches ont ainsi été effectuées sur l'identification du type de matériel impliqué dans la production sonore. Ainsi, dans [GM06], les auteurs enregistrent le son de différentes plaques suspendues par des ficelles et frappées par une bille en acier. Ils proposent aux participants d'identifier le type de plaque (acier, verre, bois ou plexiglas) à partir de l'écoute. L'expérience d'identification des différents matériaux montre de très bons résultats d'identification. Les quelques erreurs commises par les participants sont dues à la proximité acoustique des groupes acier/verre et bois/plexiglas.

Certaines études portent sur la géométrie des objets en vibration. Par exemple, dans une expérience décrite dans [KPT00], les participants identifient avec succès les formes de triangle, de rectangle et de cercle, que ces objets soient faits du même matériau ou non. L'étude met en lumière le rapport entre les dimensions des matériaux, le signal émis, et la forme perçue.

Enfin d'autre recherches portent sur le lien entre le son et l'action qui a été effectuée. Dans [WV84], les participants obtiennent des résultats très satisfaisants dans l'identification du rebond ou de la casse après une chute d'un verre. Les auteurs montrent que dans la perception de ces deux actions, les informations temporelles ont plus d'importance que les informations fréquentielles. De même lors de l'expérience du remplissage du contenant décrite précédemment (partie 5.1.2), le son permet de préciser l'aspect temporel de l'action effectuée [CP00]. Enfin, dans une autre étude, Guillaume Lemaitre montre que des participants identifient mieux l'action que le matériel lors de l'écoute de sons produits par des cylindres de différentes matières, et animés de différents mouvements [LH12].

Certaines de ces études démontrent donc les capacités des participants à identifier les propriétés de l'objet ou de l'action. Des associations incorrectes entre un son et une source sonore peuvent également être mises en évidence. Il semble par exemple que les bruits de pas lourds et sonores soient systématiquement associés avec des marcheurs masculins [Lut07].

Capacité d'identification

Ces dernières études montrent des résultats obtenus par un ensemble de participants. Toutefois, la capacité à inférer sur des propriétés de la source sonore à partir du son est subjective, et dépend considérablement de l'habitude de l'auditeur ou de son expertise. Ballas montre ainsi dans [Bal93] que l'identification des propriétés de la source sonore dépend :

- de paramètres acoustiques,
- de la fréquence écologique (fréquence à laquelle le son est entendu dans la vie quotidienne),
- de l'incertitude causale (quantité de causes différentes reportées par des participants),
- et de la typicalité du son (capacité à être un bon représentant de la catégorie de son).

À l'expertise de l'auditeur s'ajoute le rôle du contexte, qui est également très important. Certains travaux ont ainsi mis en évidence les liens cognitifs entre le traitement des sons environnementaux et celui de la parole. Dans le cadre de la parole en effet, les mots entourant un mot mal entendu permettent de le comprendre. De même une scène sonore peut nous permettre d'identifier un son [HB80]. De la même façon que les mots, la perception des sons « homonymes » pouvant être identifiés de différentes manières a également été étudiée [BM91].

En parallèle de ces recherches d'identification, d'autres analyses sont effectuées sur la façon dont les évènements sonores similaires sont regroupés. En référence aux études ci-dessus, nous pourrions par exemple considérer l'ensemble des sons produits par les mêmes matériaux, par des objets de même forme, ou par la même action.

5.1.4 Catégorisation

Théorie classique et prototypique

Le processus de regroupement selon des propriétés similaires est couramment appelé catégorisation [Gol94]. La catégorisation permet de simplifier la diversité de notre environnement, en regroupant les objets qui ont des propriétés similaires, et donc éventuellement d'appliquer un comportement semblable par rapport aux objets de la même catégorie. Une fois un objet rattaché à une catégorie, le principe d'inférence permet de déduire de la catégorie des propriétés

qui ne sont pas directement accessibles depuis l'objet. Par exemple, la vision très partielle d'une table pourrait suffire à utiliser les propriétés de la catégorie « table » et de savoir que cet objet possède quatre pieds.

La catégorisation a été introduite dans l'antiquité par Aristote qui définissait les catégories par des conditions nécessaires et suffisantes. Par exemple, dans un ensemble de formes de couleur, il peut être assez facile de déterminer le sous-ensemble des carrés rouges. Suivant cette approche, les catégories sont bien définies et ont des limites claires et stables.

Dans les années 1970, Eleanor Rosh soumit l'idée que les catégories n'étaient pas forcément décrites par un ensemble précis de règles, mais pouvaient être définies selon un processus global dans lequel les objets s'approchent au mieux d'un prototype [Ros73]. À la différence de l'approche aristocienne, les catégories ne sont pas des ensembles stables mais évoluent selon les connaissances des personnes. Dans ce cadre, sur le territoire américain, un merle est un meilleur prototype d'oiseau qu'une autruche. L'existence du prototype a depuis été remise en cause par certains chercheurs préférant la notion de membre représentatif [Dub93].

Similarité

La conception prototypique de la catégorisation conduit à utiliser la notion de similarité comme un élément de base de la création des catégories. Les catégories sont ainsi formées selon des propriétés similaires ou un « air de famille » avec le prototype [RM75]. En revanche certaines catégories peuvent être formées dans un but précis, par exemple la catégorie des « choses à vendre au vide-grenier » [Bar83]. La similarité perceptive ne suffit donc pas toujours à expliquer la formation des catégories, qui dépend également des connaissances et du contexte [Tve77]. Différentes conceptions du rôle de la similarité dans la catégorisation ont été revues dans [Gol94].

Organisation cognitive

Dans sa théorie, Rosh propose une organisation entre les catégories, qui permet des relations d'inclusion [RL78], que nous pouvons également appeler « abstractions »[Guy96]. Trois niveaux sont ainsi définis : le niveau de base, les catégories super-ordonnées et les catégories supra-ordonnées. Ainsi à la question « sur quoi êtes vous assis ? », les participants répondent plus volontiers « une chaise » (niveau de base), qu'« un meuble » (catégorie super-ordonnée) ou qu'« une chaise de bureau » (catégorie supra-ordonnée). Le niveau de base présente de plus un équilibre entre les propriétés similaires au sein de la catégorie et des différences avec les autres catégories de même niveau. Il maximise ainsi l'efficacité d'utilisation de la catégorie, ou du mot y faisant référence.

Catégorisation sonore

Nous avons vu dans le chapitre 2, que Bregman s'est appuyé sur la théorie de la forme (gestalt) pour expliquer notre perception des évènements au sein des paysages sonores (voir partie 2.1.3). Cette faculté peut être comparée à la dissociation de la forme et du fond dans une image. Par rapport à la théorie des catégories, la reconnaissance de la forme peut donc s'expliquer par un apprentissage et une organisation des connaissances en relation avec des prototypes. Dans notre perception des scènes sonores, nous serions donc constamment en train de juger si tel ou tel évènement sonore est une instance d'une catégorie.

Lien avec la classification

La catégorisation se rapproche de la classification, telle que nous l'avons vue dans le chapitre 2. Néanmoins, la catégorisation fait référence au processus cognitif alors que la classification fait référence à la tâche. Autrement dit, la classification s'attache en fait à décrire un processus fondé sur des techniques et des structures mathématiques permettant la catégorisation. Un exemple du processus de classification peut être le classement des livres d'une bibliothèque selon le nom des auteurs. La catégorisation décrit un processus plus global dont les règles ne peuvent pas toujours être précisément explicitées, par exemple l'organisation par thème des livres d'une bibliothèque.

5.1.5 Catégorisation des sons environnementaux

Expériences de catégorisation

Dans la littérature scientifique, les catégories cognitives sont mises en évidence par des expériences de classification [Gai09]. Les classes créées par un participant lors d'un test sont supposées être le reflet de catégories cognitives, utilisées dans la vie courante de manière plus ou moins consciente. Les études expérimentales s'appuient ainsi sur des tâches de classification libres et forcées pour explorer les différents modèles cognitifs de catégorisation [SSJ06]. Nous allons voir dans la partie suivante comment ce processus a été appliqué à l'écoute des sons environnementaux.

La thèse de Nancy Vanderveer représente une étape importante dans la recherche sur la perception des sons environnementaux [Van79]. Dans une expérience, elle demande à des participants d'écrire ce qu'ils entendent à l'écoute de sons environnementaux enregistrés. Les participants décrivent rarement le son lui même, mais plutôt l'action, la source sonore, et le lieu où se passe l'action. Des expériences de classification dans lequel les participants ont pour consigne de grouper les sons selon leur similarité, montre que les sons sont regroupés soit par similarités acoustiques, soit par similarités causales.

Diverses études sur la catégorisation des son environnementaux ont depuis été effectuées. Dans sa thèse, mon homonyme Frédérique Guyot reporte des expériences de catégorisation libre [Guy96]. Elle demande aux participants d'écouter 25 sons issus de la vie domestique, et de les classer selon leur similarité perceptive. Les participants doivent par la suite nommer chaque catégorie. Les catégories trouvées sont organisées selon le modèle hiérarchique décrit par Rosch. Au niveau de base, les auditeurs identifient l'action. Au niveau sous-ordonné, ils identifient la source sonore. Le niveau super-ordonné est utilisé pour décrire un mécanisme abstrait de production du son. La figure 5.2 montre l'organisation de ces catégories. Dans les expériences, très peu de sons produits par des liquides sont utilisés.

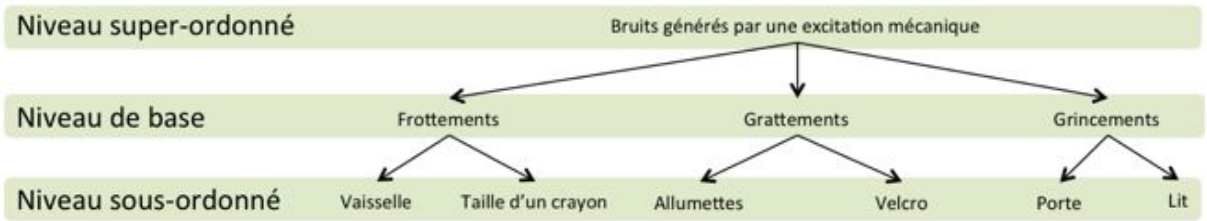


FIGURE 5.2 – Catégorisation hiérarchique des sons domestiques [Guy96].

Dans une autre étude, Marcell et ses collaborateurs proposent une expérience à partir de 120

sons et obtiennent 27 catégories [MBG+00]. Les catégories trouvées correspondent à la source sonore (animaux de ferme), à des lieux (cuisine) ou à des concepts plus abstraits (accident). La catégorie « eau/liquide » est également citée et contient six sons, dont « verser », « bulles », « rivière », « gouttes », « canalisation », et « friture ».

Dans [GKW07], les auteurs proposent une expérience de catégorisation à partir de 50 sons et trouvent 13 catégories principales. Certaines correspondent au type de source (animal, eau), d'autres catégories sont issues d'un regroupement par le contexte (sport) ou encore sur des propriétés acoustiques (hauteur fréquentielle, *pitched* en anglais).

Différents types de similarité

À la lecture de ces différentes expériences, il apparait que les participants regroupent les sons pour différentes raisons, qui ne sont pas toujours maîtrisées par les expérimentateurs ou même les participants. Il semble ainsi que les sons puissent être classés selon différents types de similarité. Ces différents types de similarité peuvent être regroupés en trois catégories :

- acoustiques (timbrale, temporelle, rythmique, etc.),
- causales (objet, action, etc.),
- sémantiques ou contextuelles (lieu, évènement, usage, etc.).

Dans [LHMS10], les auteurs s'attachent à déterminer les possibles causes influençant le type de similarité utilisée. Les participants aux expériences sont regroupés en experts, travaillant dans le domaine du son, et non-experts n'ayant pas de relation spécifique avec le son ou la musique. Dans une première expérience, l'incertitude causale du corpus de 96 sons, telle que formulée par Ballas [Bal93], est déterminée. Les participants doivent écrire pour chaque son un nom et un verbe correspondant à un objet et à une action. Un groupe de juges détermine les mots équivalents afin de calculer l'incertitude causale de chaque son. Dans une seconde expérience, les participants doivent grouper les sons en précisant quel est leur critère de similarité. Par ailleurs, un groupe de juges vérifie la similarité utilisée. Dans une troisième expérience, les participants indiquent leur taux de confiance dans l'identification des sons.

Les résultats de ces expériences permettent aux auteurs de tirer plusieurs conclusions. En général, le type de similarité utilisé n'est pas parfaitement maîtrisé par les participants. Pour les participants non-experts par exemple, la similarité acoustique est parfois utilisée alors que pour les juges, la catégorie a été effectuée sur la base d'une similarité causale. D'autre part, les participants experts utilisent plus volontiers la similarité acoustique que les participants non-experts.

Au niveau de l'incertitude causale, les sons difficilement identifiés sont en général classés selon des critères acoustiques. Les résultats montrent une grande corrélation entre la confiance des participants dans l'identification d'un son et l'incertitude causale.

Par ailleurs, nous pouvons constater que le terme le plus employé par les participants pour décrire les sons est le mot « eau ».

Autres types de catégorisation

Certains travaux récents de perception sonore s'appuient sur de l'imagerie cérébrale pour étudier le traitement cognitif des informations. Ces travaux permettent par exemple des comparaisons entre le traitement des sons environnementaux et celui du langage. Les différents traitements cognitifs mettent en évidence l'existence de catégories de haut niveau, qui peuvent être prépondérantes sur d'autres types de catégorisation.

Dans [GMM10], Bruno Giordano utilise un corpus contenant des sons environnementaux assez variés. Certains d'entre eux sont provoqués par des activités humaines (rasage par exemple), alors que certains autres sont des sons « vocaux » produits par des humains (ronflements) ou par des animaux (coassement de grenouilles). Les auteurs montrent les différences entre les processus cognitifs utilisés dans le traitement des sons provoqués par des organismes vivants et les sons produits par des objets non-vivants. Les sons d'objets non-vivants seraient ainsi traités comme des icones, ce qui les associe fortement à leur source sonore. Les sons d'organismes vivants seraient plutôt utilisés comme des symboles, et seraient associés aux référents de manière indirecte.

Les auteurs explorent de plus les différences de perception entre les sons produits par des actions et les autres sons. Le traitement cognitif est également différent pour ces deux types de sons. Toutefois, la définition de ces deux classes de sons n'est pas claire, ce qui peut soulever quelques questionnements, par exemple sur l'implication de l'humain dans la production du son. Ainsi le son « vin versé dans un verre » est marqué comme *action*, alors que « l'eau remplissant doucement un lavabo » et « chasse d'eau » sont classés dans la catégorie *non-action*. Au final, Il semblerait que la classe action représente les sons produits par des gestes. Les sons produits par des mouvements évoquant clairement un geste humain seraient donc perçus de manière différente des autres sons.

Par ailleurs, depuis la découverte des neurones miroirs [RC04], un nombre important d'études a été effectué afin de prouver des liens entre la perception d'une action et son exécution. Ainsi la perception d'un son produit par une action pourrait actionner les mêmes zones du cerveau que celles utilisées pour effectuer l'action [AP10].

5.1.6 Taxonomie des sons environnementaux

Taxonomie de Gaver

À notre connaissance, les expériences les plus proches de notre problématique de taxonomie de sons de liquides ont été effectuées très récemment à l'IRCAM par l'équipe « Perception et Design Sonores » [HLM$^+$12]. Elles visent à valider perceptivement une taxonomie des sons environnementaux selon la source physique du son, établie par Gaver au début des années 90 [Gav93].

Gaver propose ainsi des catégories de sons entendus lors de « l'écoute de tous les jours » selon la source physique à l'origine de ces sons. Il propose une classification précise des différents types d'interactions impliquées dans la production de sons. Les sons sont classés selon le type de matériel impliqué dans la production : solide, liquide ou gaz. La figure 5.3 illustre les principales catégories venant de cette classification. Cette approche est régulièrement citée aujourd'hui, par exemple pour le classement des sons dans la base de données Freesound [20] [RJK$^+$10b].

Au delà de ces trois catégories principales de la figure 5.3, Gaver décrit aussi des catégories hybrides. Ainsi, pour les sons de liquides, les sons hybrides peuvent venir d'interactions :

- liquide / solide, par exemple la pluie, le remplissage d'un contenant,
- liquide / gaz, par exemple le pétillement, ou le sifflement,
- liquide /solide / gaz, comme par exemple un bateau à moteur.

La figure 5.4 résume ainsi la taxonomie de Gaver relative aux sons de liquides, et montre des exemples de sons, en terme d'objets impliqués ou d'actions.

20. http://www.freesound.org/

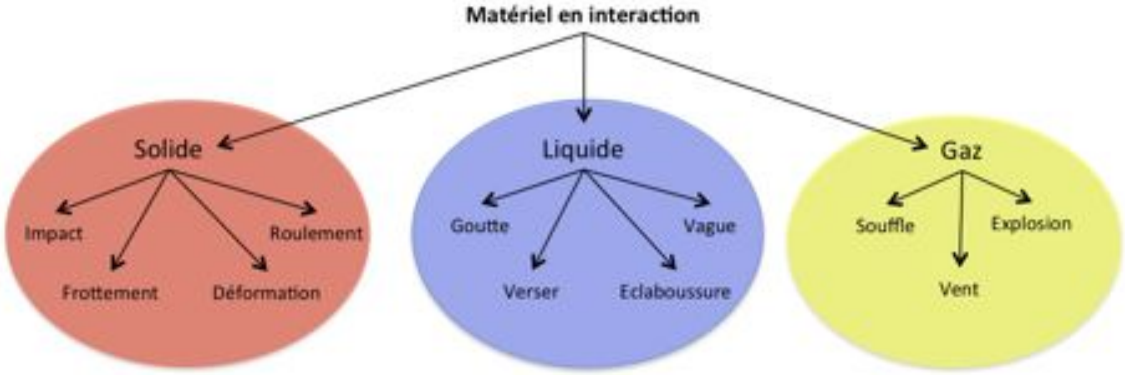


FIGURE 5.3 – Taxonomie des sons environnementaux inspirée par Gaver [Gav93, RJK+10a].

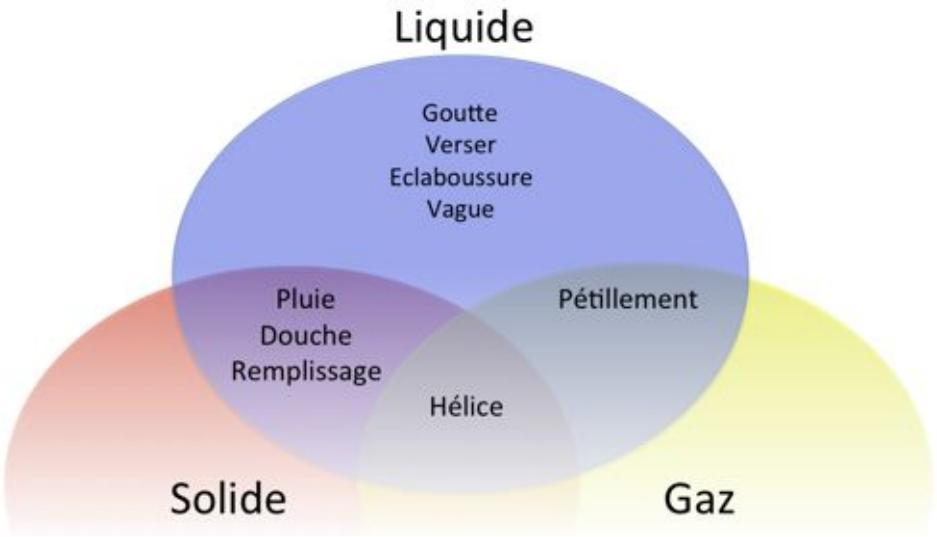


FIGURE 5.4 – Taxonomie des sons de liquides inspirée par Gaver [Gav93].

Limites

Cette étude de Gaver est très intéressante car elle fournit une taxonomie des sons environnementaux, et plus précisément des sons produits par les liquides. Néanmoins, selon l'auteur lui même, cette classification n'est pas exhaustive. De manière générale, cette taxonomie a été dictée par une analyse qualitative, et n'est pas validée par des expériences. De plus, les sons provenant de machines ou produits par la voix ne sont pas inclus.

Au niveau de la classe liquide, les différents sons d'eau que nous avons observés peuvent difficilement être associés aux quatre sous-classes (gouttes, verser, éclaboussures, et vagues, pour *drip, pour, splash* et *ripple*). Par exemple, nous ne savons pas comment classer le son produit par l'eau du robinet coulant dans un lavabo rempli. Ce son de flux d'eau tombant dans l'eau est pourtant très présent dans notre corpus d'activités quotidiennes.

Par ailleurs, selon notre modèle physique décrit au chapitre précédent, les sons de liquide

viennent principalement de la vibration de bulles d'air dans le liquide. L'utilisation de la catégorie « liquide » et de la catégorie hybride « liquide/gaz » semble donc mal adaptée à la description des sons de liquides par rapport à leur source physique.

Catégories perceptives des sons environnementaux

Dans [HLM+12], Olivier Houix et les co-auteurs interrogent la validité perceptive de la classification de Gaver. Le but de cette étude est d'observer si les classifications effectuées par les participants lors d'expériences de catégorisation libre reflètent les catégories proposées par Gaver basées sur les interactions physiques.

Les expériences sont effectuées par des participants non-experts sur un corpus de sons facilement identifiables. Dans une première expérience, 60 stimuli, représentant différents types de sons expérimentaux inclus dans la taxonomie de Gaver, sont regroupés par les participants. Après avoir classé tous les sons, les participants doivent nommer les classes constituées.

Les principales classes trouvées à l'issue de ces expériences sont similaires aux grandes catégories proposées par Gaver : solide, liquide, gaz. Par rapport à la classification de Gaver, la catégorie « machine » apparait également. Toutefois, les sous-catégories des quatre groupes (solide, liquide, gaz et machine) sont difficilement analysables.

Les résultats montrent également que les sons dont la production est en partie due à l'élément liquide sont systématiquement classés dans la catégorie « liquide ». Cette observation est en contradiction avec la distinction de la taxonomie de Gaver entre la classe « liquide » et les classes hybrides comme « liquide/air » ou « liquide/solide ».

Dans une deuxième expérience, les auteurs s'intéressent spécifiquement au son produit par les solides pour étudier les sous-catégories de cette classe. Cette fois, les résultats s'éloignent des sous-catégories de la classe « solide » proposées par Gaver. Les analyses montrent deux catégories principales, appelées « interactions discrètes » et « interactions continues ». La figure 5.5 présente ainsi les différentes catégories impliquées dans cette étude de sons de solides.

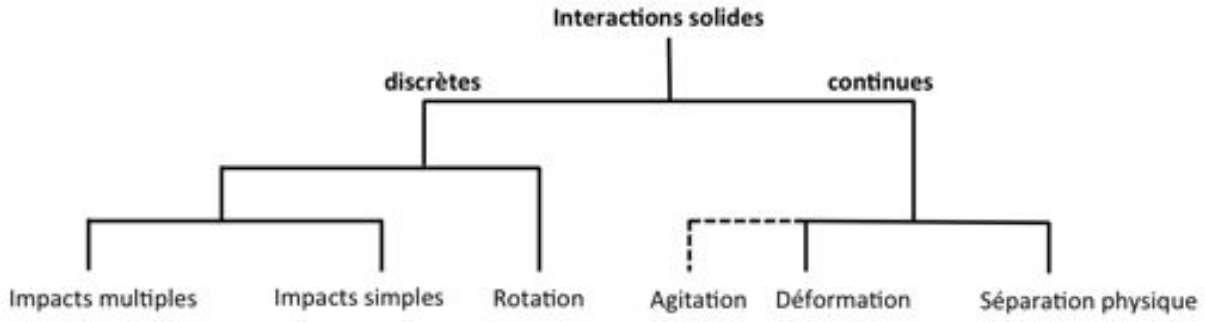


FIGURE 5.5 – Catégories de sons de solides [HLM+12].

Les résultats de ces expériences sont très intéressants et permettent d'utiliser une taxonomie des sons environnementaux validée par des expériences perceptives. Elles proposent une organisation générale de la perception des sons environnementaux, et une description plus précise des catégories de sons de solides. De plus, les résultats sont cohérents avec la catégorisation hiérarchique présentée dans [Guy96] et [Dub00] (voir figure 5.2), car dans cette dernière étude l'identification de la source sonore apparait à un niveau inférieur.

Par contre, si les sous-catégories relatives à l'élément « solide » ont été étudiées, aucune expérience spécifique n'a été effectuée sur les sons de la catégorie « liquide ». Cette catégorie nécessite donc des expériences supplémentaires pour pouvoir déterminer les différents groupes de sons de liquides.

Au cours d'un séjour à l'IRCAM, j'ai réalisé avec toute l'équipe « Perception et Design Sonores » et plus particulièrement Olivier Houix une série d'expériences que nous allons maintenant décrire.

5.2 Spécification des expériences

5.2.1 Motivations

Le but de cette étude est d'obtenir des catégories de sons de liquides validées perceptivement. Nous allons effectuer une expérience de catégorisation libre sur un corpus de sons de liquides. Nous espérons ainsi que les classes créées par les participants lors cette expérience reflètent les catégories cognitives mises en jeu dans l'écoute et l'identification des sons de liquide.

Pour que l'expérience puisse refléter ces catégories, il est important de maîtriser au mieux le type de similarité utilisé par les participants lors de la classification. Nous avons en effet vu dans l'état de l'art que plusieurs types d'écoute [Gav93], et plusieurs types de similarité pouvaient être utilisés pour classer les sons : acoustique, causale, et contextuelle [LHMS10].

Nous ne cherchons pas ici à établir des catégories acoustiques (par exemple *son court, son grave*), ni des catégories contextuelles (par exemple *cuisine, ménage*), mais des catégories reflétant le phénomène physique à la base de la production du son (par exemple *remplissage d'un récipient*).

5.2.2 Spécification du corpus

Sons du domicile

Par souci de cohérence avec nos précédents travaux, nous allons effectuer notre expérience de catégorisation sur des sons pouvant être entendus au sein du domicile. Cette restriction permet de plus d'assurer une continuité avec les études précédentes de l'IRCAM effectuées sur les sons de solide [HLM$^+$12].

Liquides et hybrides

Nous devons nous assurer que les catégories trouvées à l'issue de l'expérience soient représentatives de l'ensemble des sons de liquide du quotidien. Les expériences de l'état de l'art montrent en effet que les sons hybrides (liquide/solide, liquide/gaz, ou liquide/gaz/solide) sont classés par les participants dans la catégorie liquide [HLM$^+$12]. Il semble important d'utiliser un ensemble de sons de liquide le plus varié possible, tout en conservant la contrainte que ces sons puissent être entendus dans la vie quotidienne au sein de son domicile.

Identifiabilité et expertise

Nous avons vu dans l'état de l'art que les sons mal identifiés étaient généralement classés selon leur similarité acoustique [LHMS10]. Pour privilégier la similarité causale, le corpus devra être constitué de sons facilement identifiables. Les sons difficilement identifiables seront donc supprimés par les participants lors d'une expérience préalable.

De même, nous limiterons les participants de l'expérience de catégorisation à des personnes non expertes en acoustique ou en musique, ces derniers pouvant privilégier la similarité acoustique [LHMS10].

Durée et sonie

De manière générale, il est important de ne pas constituer de groupes de sons caractérisés par des paramètres similaires indépendants de leur cause. Ces sons pourraient en effet être classés ensemble en fonction de ces paramètres, et non de leur cause. Pour homogénéiser le corpus nous allons modifier deux de ces paramètres : la durée et la sonie.

La durée des sons peut être un paramètre sur lequel les participants pourraient s'appuyer pour former des classes, par exemple la classe des « sons longs ». Le corpus doit donc présenter une certaine homogénéité dans la durée des sons. De plus, pour que les sons puissent être écoutés plusieurs fois dans l'expérience, leur durée doit être relativement courte. Nous allons donc, dans une étape préalable d'édition, homogénéiser les durées des sons.

La sonie des sons enregistrés dépend de la distance entre le microphone et la source sonore. Elle est également liée à l'énergie du signal, et peut être modifiée par un traitement sur le signal. Une modification en sonie, dite « écologique », sera donc effectuée par des participants dans une première expérience afin d'homogénéiser la distance de la source sonore perçue.

Sons humains et sons d'action

Selon des études précédentes, les sons provoqués par des bruits humains ou par des actions provoquent un traitement cognitif différent [GMM10]. L'utilisation de ce type de sons dans notre expérience de catégorisation pourrait mener à l'obtention de classes spécifiques et masquer les catégories liées à l'évènement physique à la base du son. Les sons de bruits humains, par exemple le son produit par un gargarisme, ne seront donc pas utilisés dans notre corpus.

Au niveau des actions, il semble difficile de définir précisément les sons produits par les actions humaines. Certains sons semblent identifiables en tant que son d'action parce qu'ils sont constitués d'une séquence sonore que l'on assimile à une production humaine. Nous pouvons considérer par exemple, la séquence suivante :

ouverture d'une cannette - versement dans un verre avec glaçons - pétillement

Cette séquence peut être facilement assimilée à une intervention humaine, alors que l'écoute d'une partie de la séquence seulement le sera moins directement. Par contre, l'écoute d'une partie de la séquence rend beaucoup plus difficile l'identifiabilité du son. Nous avons donc du trouver un compromis entre sons identifiables et sons d'actions.

D'autres sons ayant pour origine des gestes ou des mouvements sont également identifiables comme sons d'actions. C'est le cas par exemple d'un pommeau de douche agité, ou d'un récipient secoué. Toutefois, pour privilégier la variété du corpus, nous avons conservé ce type de sons.

5.3 Constitution du corpus

5.3.1 Inventaire des lexèmes liés au son d'eau

Pour constituer ce corpus nous avons identifié de manière lexicale les sons faisant intervenir l'élément liquide dans leur production. Dans cette approche, nous avons travaillé à la fois sur les langues française et anglaise. Pour identifier les lexèmes liés aux sons de liquide, nous avons utilisé plusieurs dispositifs :

Taxonomie de Gaver Une liste de base a été constituée à partir de la taxonomie de Gaver. Cette étude nous a donné plusieurs types d'actions et d'exemples sons, qui selon Gaver sont représentatifs des sons :

- liquides (couler, verser, éclabousser, clapotis),
- hybrides (remplir, pluie).

Questionnaires Cette première liste de mots a été complétée par l'utilisation de questionnaires, contenant la question suivante :

- *Dressez une liste de tous les événements sonores faisant intervenir dans la production du son un élément liquide (matière plus ou moins visqueuse), et que vous pourriez entendre à l'intérieur d'une maison.*
- *Notez si possible une brève description de chaque événement sonore (lieu, objet(s), action...).*

Quatre participants, travaillant à l'IRCAM ou à l'IRIT, ont répondu à ce questionnaire. De nombreux lexèmes ont été ajoutés à la liste comme par exemple le verbe *vaporiser*.

Réseau de mots Nous avons utilisé des réseaux de mots, notamment le projet *Wordnet*[21] dans lequel est défini un ensemble de relations entre les mots. Par exemple à partir du mot voiture, nous pourrions obtenir :

- des relations principales ou synonymes (automobile),
- des relations super-ordonnées (véhicule),
- des relations sous-ordonnées (berline),
- des relations de méronymie liant un tout à une partie (pare-brise).

Ce réseau de mots peut également contenir d'autres relations, comme l'antonymie. La figure 5.6 nous montre les relations proposées pour le mot *liquid*. Elle est issue du site *Wordvis*[22] qui permet de visualiser les liens entre mots du réseau *Wordnet*. Les disques correspondent à différentes définitions du terme *liquid*, qu'il soit utilisé comme un nom (en rouge) ou comme un adjectif (en jaune). Les liens *sim* signifient « est similaire à » et les liens *is* signifient « est un type de ».

Base de données L'exploration des bases de données sonores nous a conduits à considérer des sons qui n'ont pas été produits par les méthodes précédentes : par exemple *récipient secoué*.

5.3.2 Collectage de fichiers audio

Dans des bases de données sonores, nous avons sélectionné tous les sons dont le label contenait un des lexèmes trouvés dans l'étape précédente. Nous avons ainsi collecté un ensemble de plus de 700 fichiers audio à partir des bases de sons suivantes :

- Auditory Lab[23]
- BBC Sound Effects Library - Original Series[24],
- Blue Box Audio Wav[25],
- Hollywood Edge Premiere Edition I, II and III[26],

21. http://wordnet.princeton.edu/
22. http://wordvis.com/
23. http://www.psy.cmu.edu/ auditorylab/website/index/home.html
24. http://www.sound-ideas.com/sound-effects/bbc-1-40-cds-sound-effects-library.html
25. Best Service GmbH, München, Germany
26. The Hollywwod Edge, Hollywood, USA

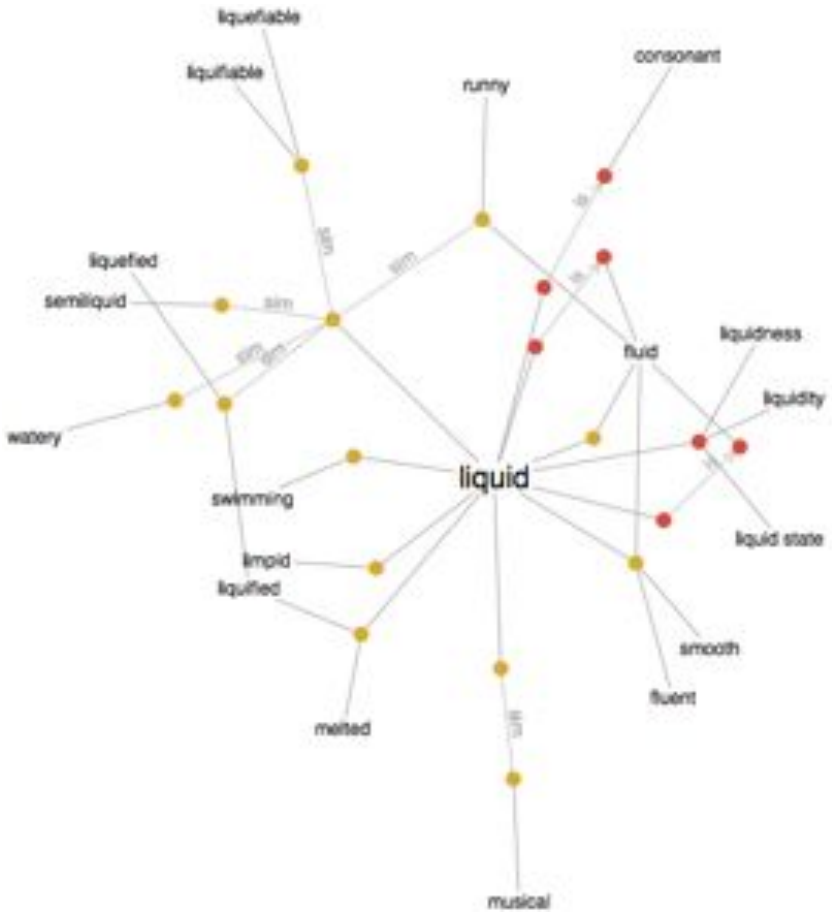


FIGURE 5.6 – Liens entre mots du réseau *Wordnet*. Figure extraite du site *Wordvis*.

– Sound Ideas General Series 6000 [27],
– SoundScan v2 Vol. 61 [28].

La plupart de ces bases de sons sont disponibles dans le commerce, et sont destinées à la création sonore. La base de données créée dans le laboratoire *Auditory Lab* est légèrement différente, puisqu'elle est dédiée à la recherche et organisée selon la taxonomie de Gaver. Elle propose des sons enregistrés en laboratoire, et classés selon les thèmes liquides, solides et gaz.

5.3.3 Élimination des redondances

Concrètement, notre ensemble de sons est beaucoup trop important pour pouvoir être écouté par des participants lors d'une expérience. De plus, il existe de nombreuses redondances inutiles. Nous avons donc effectué un traitement sur le corpus pour éliminer les redondances trop évidentes, c'est-à-dire les sons acoustiquement similaires et produits par la même cause.

Pour identifier précisément les redondances, nous avons trié les sons en constituant une « carte sonore ». Cette carte fut créée à l'aide du logiciel VUE [29], permettant de positionner les sons

27. Sound Ideas, Ontario, Canada
28. Ultimate Sound Bank, Paris, France
29. http://vue.tufts.edu/

dans un espace à deux dimensions, de les grouper et de les écouter.

Pour disposer les sons dans un plan, nous avons utilisé les classes hybrides de Gaver : *liquide, solide, gaz*, auxquelles nous avons ajouté la classe *visqueux*. L'écoute des sons nous a amenés à utiliser d'autres critères, par exemple le débit d'eau. La figure 5.7 illustre les groupes de sons utilisés dans ce processus. Les étiquettes utilisées pour ces groupes sont arbitraires, et servent uniquement à les différencier.

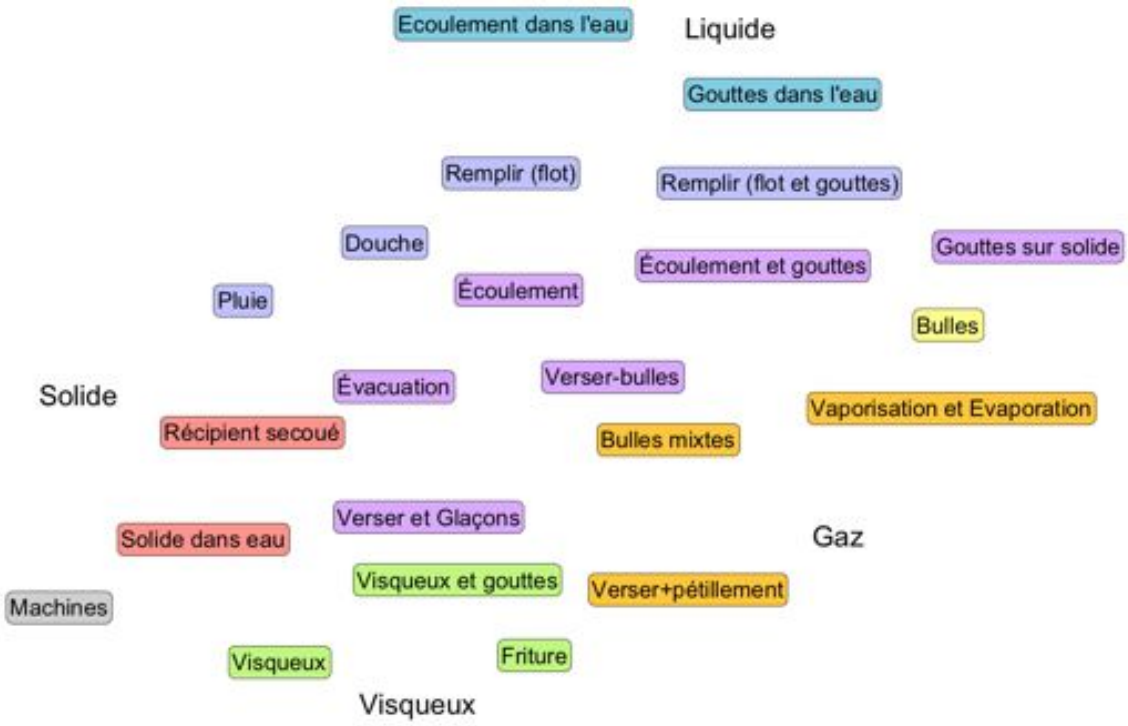


FIGURE 5.7 – Carte sonore de sons de liquide.

Une fois les sons classés par redondance, nous avons diminué la taille de notre ensemble en supprimant des sons dans les groupes de cardinalité élevée. Cette opération a permis d'éliminer une partie importante des sons collectés pour obtenir un corpus effectivement utilisable pour la réalisation d'une expérience. À l'issue de ce processus, 81 stimuli sonores différents ont été conservés.

La liste des sons est fournie en annexe E. Chaque son est identifié par un label fourni par les librairies sonores qui a été traduit en français. L'ensemble des labels a été homogénéisé en terme de détails.

5.3.4 Édition des sons

Dans cette étape d'édition, nous avons homogénéisé la durée des sons. La figure 5.8 montre le résultat de cette édition, en terme de durée de son croissante.

Au final, les caractéristiques temporelles du corpus sont les suivantes :

- Durée moyenne = 3 s,
- Durée minimale = 0,4 s,
- Durée maximale = 5,4 s.

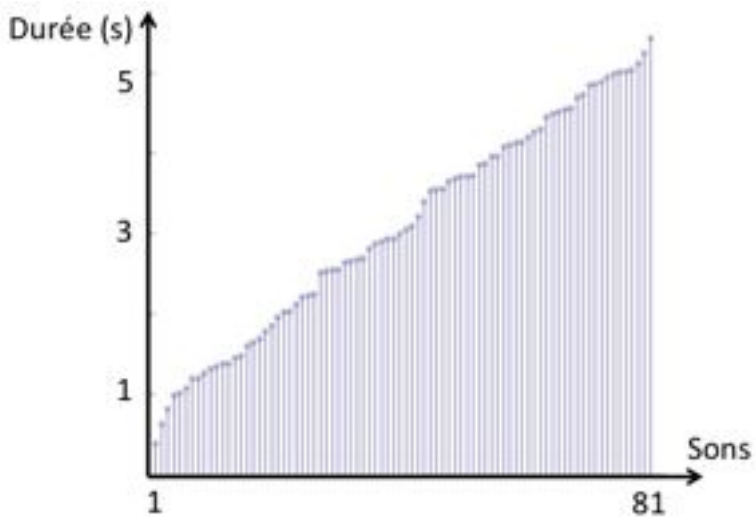


FIGURE 5.8 – Sons du corpus classés par durée croissante.

5.4 Description des expériences

Nous allons présenter dans cette section l'expérience de catégorisation libre. Comme nous l'avons expliqué précédemment, deux expériences préalables sont nécessaires à la préparation du corpus. Nous présentons dans cette partie la première expérience qui a pour but l'égalisation écologique, et la deuxième expérience d'identification des sons. Nous terminerons par l'expérience de catégorisation libre. Les résultats de cette dernière expérience seront présentés dans la partie suivante.

Toutes les expériences ont été effectuées à l'IRCAM en cabine d'écoute insonorisée. Au niveau matériel, les stimuli sonores ont été joués par l'intermédiaire d'un ordinateur MacBookPro, d'une carte son RME 400 et d'enceintes YAMAHA MSP5.

Pour chaque expérience et pour chaque participant, les sons sont présentés dans un ordre différent, établi de manière aléatoire. Chaque participant a effectué une et une seule des trois expériences.

5.4.1 Expérience préparatoire 1 : Égalisation écologique

Présentation

L'objectif de cette expérience est de modifier la sonie des éléments du corpus pour constituer un ensemble homogène et réaliste. En effet, comme les sons proviennent de bases de données différentes, les conditions d'enregistrement, et les post-traitements peuvent être différents. Par exemple, un enregistrement de goutte d'eau pourrait être joué très fort, alors que ce son est d'intensité plutôt faible lorsque nous l'entendons dans la vie réelle. Des différences trop importantes avec la sonie de la vie réelle peuvent rendre les sons moins réalistes, et l'identification de la cause sonore moins évidente, ce qui pourrait privilégier d'autres types de similarités.

Calibrage

Dans ces expériences, nous utilisons un son de référence, un son de robinet ouvert, qui est joué avant chaque nouveau stimulus. Pour calibrer le volume d'écoute de ce son de référence, nous avons effectué des mesures de l'intensité sonore d'un vrai son de robinet. Nous avons utilisé un sonomètre pour calculer la pression acoustique d'un flux d'eau tombant dans l'eau. Nous obtenons une valeur de pression acoustique de 64,5 dBa en crêtes. À partir de cette valeur nous pouvons calibrer le volume sonore de notre système d'écoute pour que la pression acoustique en cabine soit équivalente à celle enregistrée devant le lavabo. Le sonomètre a été placé en cabine à la position d'un auditeur assis. La distance entre le sonomètre et les enceintes est égale à celle mesurée lors du calibrage entre le sonomètre et le flux d'eau (voir figure 5.4.1.0).

FIGURE 5.9 – Calibrage devant un robinet ouvert et en cabine d'écoute.

Implémentation

L'expérience fut implémentée grâce au logiciel PsiExp [30]. Ce langage de programmation dérivé du langage C permet de contrôler les stimuli, les interfaces graphiques, et d'enregistrer les résultats des participants. Au sein de ce logiciel, les stimuli sont joués par le logiciel MaxMsp [31].

Procédure

Les participants doivent écouter des paires de sons constituées du son de référence et d'un autre son du corpus. Leur tâche est d'ajuster le deuxième son pour qu'il corresponde au mieux au niveau perçu dans un contexte naturel d'écoute au sein de l'habitation. La consigne de l'expérience est disponible en annexe F. Les participants peuvent écouter chaque paire de sons plusieurs fois. Les stimuli sont présentés avec leur label, de façon à ce qu'ils soient identifiés par les participants. Un curseur permet alors de changer le volume sonore du second son (voir figure 5.10).

Résultats

Neuf participants travaillant à l'IRCAM ont effectué cette expérience. Les résultats de cette expérience sont présentés en annexe G. Pour certains sons, le consensus entre participants est

30. http://mt.music.mcgill.ca/~wwwmpcl/docs/psiexp/psiexpHtml11apr/psiexp.html
31. http://cycling74.com/

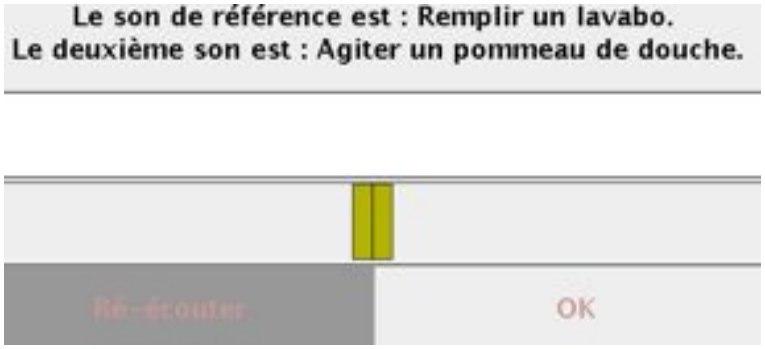


FIGURE 5.10 – Interface de l'expérience d'égalisation sonore.

assez important, par exemple le son « Agiter un verre contenant des glaçons ». D'autres sons, comme les sons de fritures, présentent des différences très importantes.

À l'issue de cette expérience le volume de chaque son du corpus a été modifié en fonction du volume sonore médian choisi par les participants.

5.4.2 Expérience préparatoire 2 : Identification

Présentation

Le but de l'expérience d'identification est d'éliminer les sons mal identifiés, pour éviter qu'ils ne soient classés selon des critères acoustiques. Selon les études précédentes, la confiance des participants dans leur identification du son est fortement corrélée avec leur capacité à effectivement identifier le son [LHMS10]. Nous utilisons donc une mesure de confiance dans l'identification. Cette mesure est bien moins coûteuse en temps qu'une expérience réelle d'identification qui nécessiterait une analyse sémantique des résultats. Cette étape permet d'éliminer les sons pour lesquels la majorité des auditeurs ne sont pas confiants dans l'identification du son. Néanmoins, les sons présentant un indice de confiance élevé ne sont pas forcément identifiés de la même façon par les participants.

Procédure

Cette expérience a également été implémentée dans le logiciel PsiExp. La consigne complète est présentée dans l'annexe F. Dans un premier temps, les participants doivent écouter tous les sons. Ils doivent ensuite écouter chaque son (deux écoutes possibles) et répondre à la question suivante :

Parvenez vous à vous représenter la cause du son ?

Ils doivent alors choisir l'une des 5 réponses suivantes :
- *Je ne sais pas du tout*
- *Je ne suis vraiment pas certain*
- *J'hésite entre plusieurs types de causes*
- *Je suis presque certain*
- *Je me représente parfaitement la cause*

Résultats

Nous avons effectué cette expérience avec 13 participants travaillant à l'IRCAM. La figure 5.11 présente les résultats des participants. Pour chaque son, la figure représente la moyenne des résultats (points rouges) et l'écart type (en bleu, de chaque coté de la moyenne).

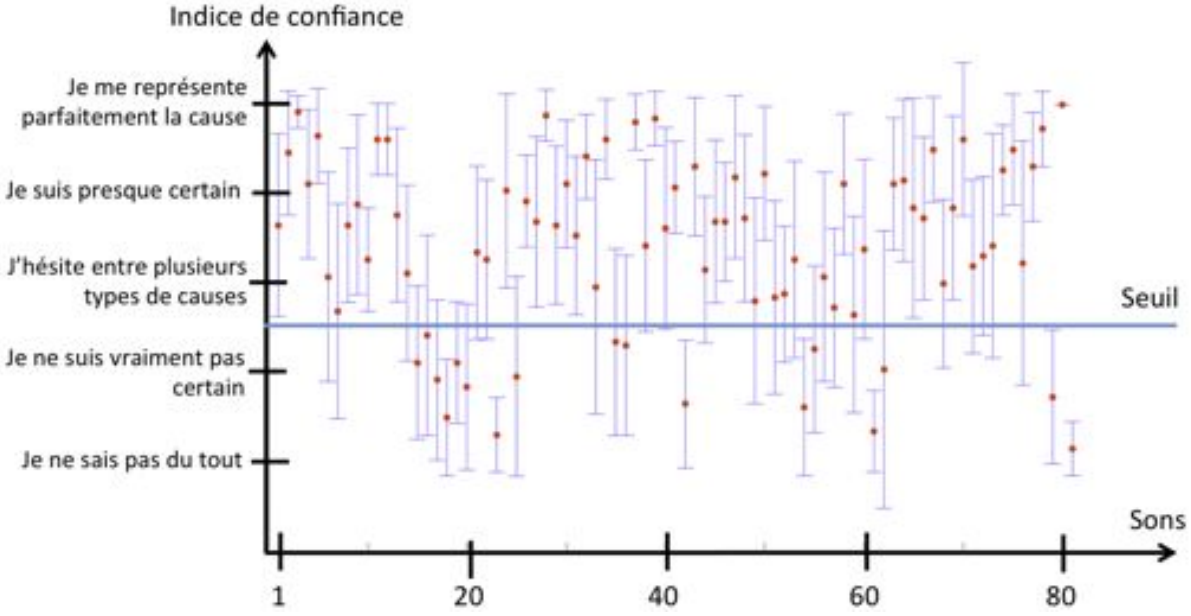


FIGURE 5.11 – Identification des sons sur une échelle allant de 0 à 4.

Nous avons utilisé un seuil pour éliminer les sons, placé sur la figure à 1,5 soit entre les réponses *Je ne suis vraiment pas certain* et *J'hésite entre plusieurs types de causes*. À l'issue du traitement, 17 sons ont été éliminés, dont le son *goutte d'eau sur une plaque chaude* et le son *éclaboussures sur du carton*. Le nombre de sons supprimés résulte, une fois encore, d'un compromis entre la variété des sons du corpus et de leur identifiabilité.

5.4.3 Catégorisation libre

Présentation

La dernière expérience constitue l'expérience principale. C'est une expérience de catégorisation libre, le nombre de classes de sons n'étant pas fixé. Le but de cette expérience est d'obtenir les classes de sons de liquides correspondant aux catégories perceptives. Dans l'obtention de ces classes, nous avons essayé de favoriser au maximum la similarité causale.

Implémentation

Pour cette expérience, nous avons utilisé le logiciel TCL-LabX [32]. Ce logiciel est développé par Pascal Gaillard du Laboratoire Octogone de L'Université Toulouse 2 Le Mirail. Ce logiciel permet d'effectuer ce type d'expérience, d'enregistrer chaque action des candidats, et d'exporter les résultats dans différents formats.

32. http://petra.univ-tlse2.fr/tcl-labx/

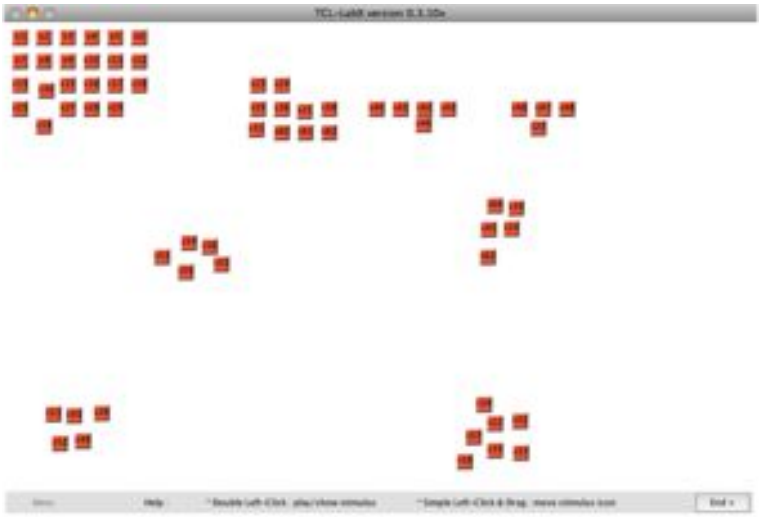


FIGURE 5.12 – Interface du logiciel TCL-LabX.

Procédure

Dans un premier temps, les participants ont pour consigne d'écouter tous les sons. L'expérience se déroule ensuite en deux phases. La première phase consiste à constituer des groupes de sons. Pour cela, les participants déplacent les sons représentés par des carrés (voir figure 5.12). Ils peuvent écouter chaque son autant de fois que nécessaire. De même, il n'y a aucune restriction sur le nombre de groupes à constituer. Voici un extrait de la consigne (disponible en annexe F) qui leur est présentée :

Former des classes de sons en fonction de l'événement physique qui a causé le son. Par exemple pour les sons : « pétard », « ballon percé », « pneu de voiture qui éclate », l'événement physique qui a produit le son pourrait être « explosion ».

Dans la deuxième phase de l'expérience, les participants ne peuvent plus déplacer les sons ou modifier les groupes. Ils doivent décrire chaque classe par un mot ou une phrase.

Nous avons choisi des participants extérieurs, n'ayant pas d'expertise particulière dans le son ou dans la musique. Certains des participants proviennent d'une base de données de personnes volontaires pour passer des expériences à l'IRCAM. D'autres personnes ont été contactées par des annonces sur le site du Relai d'Information sur les Sciences de la Cognition (RISC) [33]. Ces participants ont été rétribués pour le travail effectué. 30 participants ont réalisé cette expérience.

5.5 Analyse des résultats

Les participants ont mis en moyenne 26 minutes pour effectuer la première tâche de regroupement. Ils ont effectué en moyenne 10 classes de sons. Deux personnes ont effectué le minimum de 3 classes, le maximum étant de 18 classes.

Malgré ces nombres de classes différentes, nous espérons dans ce type d'expérience que les classes formées par les participants soient équivalentes. Le découpage en un nombre de classes

33. http://expesciences.risc.cnrs.fr/

important peut en effet faire apparaître les mêmes catégories, exprimées à une granularité plus fine, qu'un découpage en quelques classes.

Toutefois, il est possible que des stratégies différentes soient utilisées par les participants. Notre méthode d'analyse consiste au final à observer les classes communes entre participants si elles existent. Il est important d'identifier au préalable des différences de stratégies individuelles pour pouvoir éventuellement traiter les résultats séparément.

5.5.1 Analyse inter-participant

Pour calculer la similarité entre les stratégies des participants, nous effectuons un traitement analogue à celui décrit dans [AVCC07]. Pour chaque participant p nous calculons une matrice de distance $D_{[p]}$. Cette matrice est une matrice carrée symétrique de la taille du nombre de sons à classer. La distance entre deux sons est 0 s'ils ont été classés ensemble, 1 sinon.

Nous utilisons le coefficient Rv [Esc73], qui fournit une valeur de similarité entre des matrices carrés symétriques. Le coefficient Rv est calculé entre chaque paire de participants. Son calcul nécessite plusieurs étapes détaillées et interprétées dans [AVCC07].

Le calcul du coefficient Rv pour chaque paire de participants produit une matrice de similarité interparticipant. C'est une matrice carré symétrique de la taille du nombre de participants. Dans cette matrice, les lignes de coefficients Rv faibles indiquent que le participant correspondant a eu un comportement singulier. Pour interpréter ces valeurs plus formellement, nous effectuons une Analyse en Composantes Principales (ACP) de cette matrice de similarité interparticipant. La figure 5.13 illustre la première (axe vertical) et la deuxième composante (axe horizontal) de cette ACP qui contiennent respectivement 35% et 5% de pourcentage de variance expliquée.

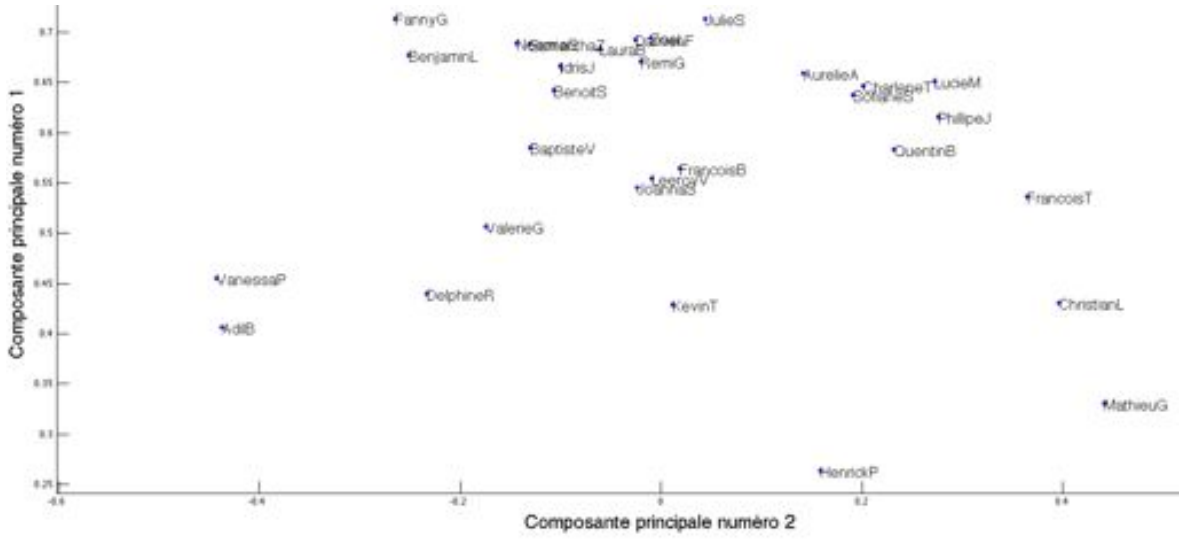


FIGURE 5.13 – Analyse en composantes principales sur le coefficient Rv.

Nous pouvons observer sur cette figure qu'un participant se détache particulièrement du groupe. Il s'agit du participant *HenrickP*, dont la projection est la plus faible des participants sur l'axe de la composante principale numéro 1. Au regard des résultats de la classification, ce participant a en fait effectué trois classes dont les verbalisations correspondent à « Sons plutôt entendus dans la cuisine », « Sons plutôt entendus dans les WC », et « Sons plutôt entendus dans

une salle de bain ». Ces verbalisations montrent que le participant a effectué sa classification sur l'unique base du contexte de production du son. Cette similarité contextuelle correspond à l'un des trois types de similarité décrit plus haut, mais ne correspond pas à la classification demandée dans la consigne, basée sur l'évènement physique qui a causé le son. Nous avons donc supprimé ce participant de notre analyse.

Nous pouvons de plus nous interroger sur la pertinence de la conservation d'autres participants quelque peu isolés, comme *MathieuG*. Plusieurs autres analyses ont été effectuées : par exemple l'analyse des 19 participants les plus similaires. Ces analyses n'ont pas mis en évidence d'autres types de stratégies différentes. Ainsi, nous présentons dans la suite une analyse unique des 29 participants restants.

5.5.2 Classification hiérarchique

Matrice de distance

Les matrices de distance individuelles sont additionnées. Les coefficients de la matrice résultat sont divisés par le nombre de participants. Nous obtenons ainsi une valeur de dissimilarité comprise entre 0 et 1 pour chaque paire de sons. Pour trouver les classes le plus souvent utilisées à partir de cette matrice, différentes méthodes peuvent être utilisées. Nous allons effectuer une analyse par regroupement hiérarchique ascendant que nous visualiserons par l'intermédiaire d'un dendrogramme [Bre93].

Choix du calcul de la distance

Nous utilisons un algorithme itératif permettant de regrouper les stimuli [Bre93]. Au début de l'algorithme chaque groupe correspond à un stimulus. Il existe donc une mesure de distance, ou dissimilarité, entre chaque paire de groupes. À chaque itération de l'algorithme, les groupes les plus proches sont fusionnés. Nous calculons alors une nouvelle mesure de dissimilarité entre le groupe issu de la fusion et les groupes restants. Cette fusion de groupes est réitérée jusqu'à l'obtention d'un groupe unique contenant tous les stimuli.

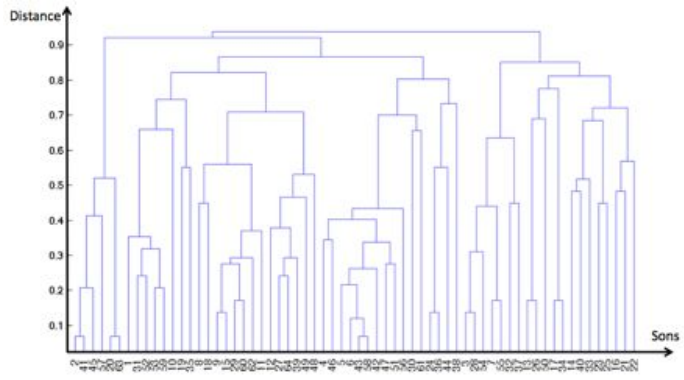


FIGURE 5.14 – Dendrogramme sur le corpus de 64 sons par la méthode WPGMA.

Le dendrogramme permet de visualiser ces groupes. La hauteur d'un lien dans l'arbre représente la distance du fusion de deux groupes. Cette valeur est appelée distance cophénétique. Ainsi la hauteur d'un lien entre deux feuilles de l'arbre, soit deux stimuli, est égale à la valeur

correspondante dans la matrice de distance. Par contre, la hauteur de fusion d'un groupe de plus de deux stimuli est une valeur approchée qui dépend des distances entre chaque stimuli.

Pour savoir si l'arbre représente au mieux les données et la matrice de distance, nous pouvons donc comparer la hauteur d'un lien dans l'arbre à la distance entre chaque paire de stimuli des sous-branches. Nous utilisons pour cela le coefficient de corrélation cophénétique qui vaut 1 si l'arbre représente parfaitement les données.

Pour un même jeu de données, le coefficient de corrélation cophénétique varie en fonction de la méthode de calcul de distance utilisée. Nous avons comparé différentes méthodes. La distance entre deux groupes est calculée en fonction de la distance des éléments de groupe différents, et peut être attribué à :

- la distance la plus petite,
- la distance la plus grande,
- la distance médiane,
- la distance moyenne appelée *WPGMA (Weighted Pair Group Method with Averaging)*,
- la distance moyenne pondérée par le nombre d'éléments des groupes, appelée *UPGMA (Unweighted Pair Group Method with Averaging).*

L'implémentation de ces différentes méthodes est présentée dans [GM07]. La méthode de la moyenne pondérée semble plus raffinée, car chaque nouvelle distance est pondérée par le nombre d'éléments dans le groupe. C'est pourtant avec une méthode plus simple, la méthode moyenne, que nous obtenons le meilleur coefficient de corrélation cophénétique, d'une valeur égale à 0,86. C'est donc cette méthode que nous choisissons, car elle semble mieux adaptée à nos données.

Choix du mode de regroupement

À partir de cette classification hiérarchique, il nous faut établir les classes de sons les plus représentatives. En termes familiers, il faut donc « couper les branches de l'arbre au bon endroit ». Une méthode courante consiste à utiliser un seuil sur la hauteur $h\{lien\}$ des liens de fusion dans l'arbre. La figure 5.15 illustre le résultat d'un seuillage.

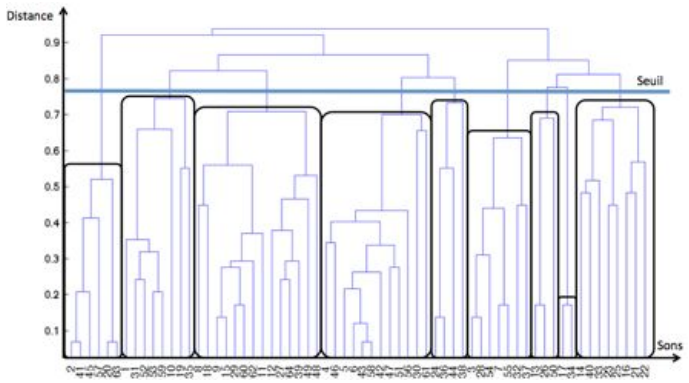


FIGURE 5.15 – Exemple de seuillage sur le dendrogramme.

Cette méthode nous permet d'obtenir des groupes de sons. Toutefois, il est assez difficile sur ce dendrogramme d'établir un seuil satisfaisant. En effet des petites variations de seuil vont modifier les groupes, qui semblent du coup assez arbitraires.

Une autre méthode de regroupement s'appuie sur le critère de compacité ou *inconsistency*

coefficient [JD88]. La compacité est calculée en comparant la hauteur $h\{lien\}$ de chaque lien de l'arbre avec la hauteur $h\{sousliens\}$ des liens inférieurs issus de la même branche.

Les liens entre deux feuilles ont un critère de compacité égal à 0. Le critère de compacité $comp\{lien\}$ des autres liens est calculé par rapport à la hauteur du lien $h\{lien\}$ et des sous-liens, ainsi que leur hauteur moyenne (moy) et leur écart type (σ).

$$comp\{lien\} = \frac{h\{lien\} - moy(h\{lien, sousliens\})}{\sigma(h\{lien, sousliens\})} \tag{5.1}$$

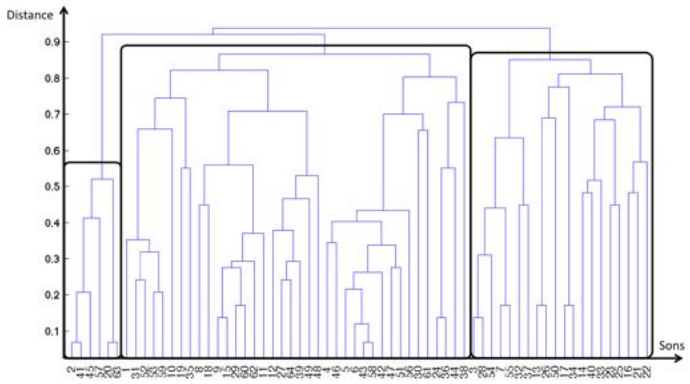


FIGURE 5.16 – Application du critère de compacité sur le dendrogramme : 3 classes sont obtenues pour un seuil de 2,1.

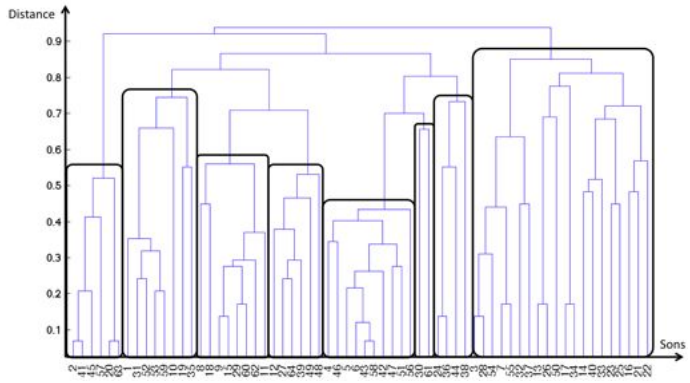


FIGURE 5.17 – Application du critère de compacité sur le dendrogramme : 3 classes sont obtenues pour un seuil de 1,7.

En utilisant un seuil sur ce critère de compacité, nous pouvons établir des groupes de hauteurs différentes dans l'arbre. La figure 5.16 illustre le résultat de ce regroupement par seuillage sur le critère de compacité. En changeant le seuil sur le critère de compacité nous pouvons obtenir un découpage de granularité différente. Ce découpage est illustré à la figure 5.17 avec 8 classes. L'utilisation du critère de compacité permet d'obtenir des ensembles plus homogènes que les ensembles obtenus à partir d'un simple seuil. Le groupe de droite est ainsi réuni dans une grande classe.

Nous allons considérer par la suite les deux découpages présentés : le découpage en 3 classes et le découpage en 8 classes. Deux classes sont communes à ces deux découpages. Nous proposons donc les labels suivants : A, B1, B2, B3 B4 B5, B6, C (voir figure 5.18).

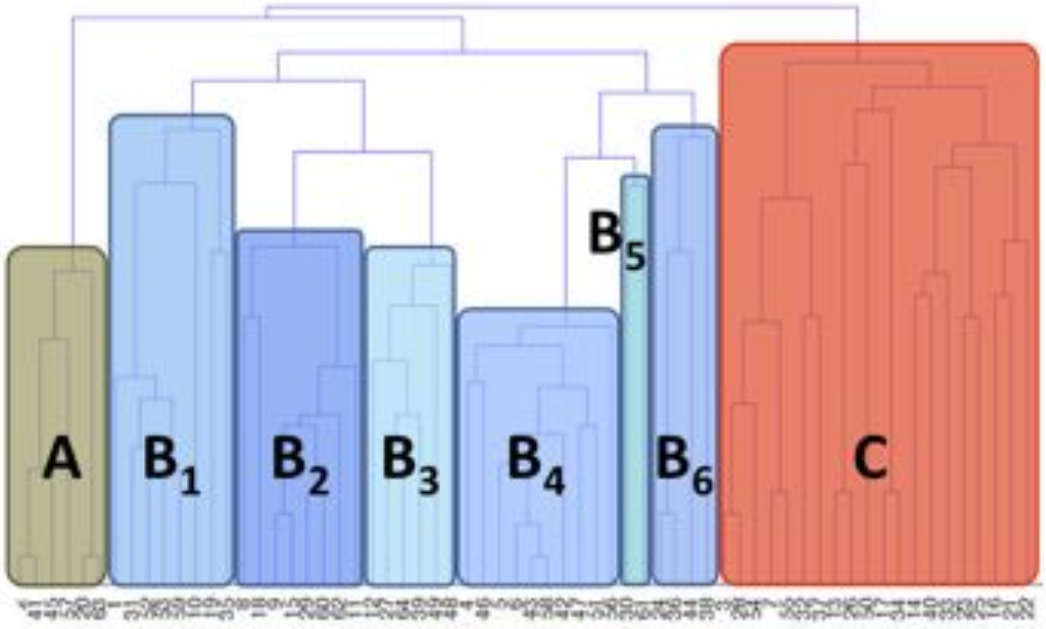


FIGURE 5.18 – Identification des classes sur le dendrogramme.

L'étape suivante de notre analyse consiste à remplacer ces labels arbitraires par des labels sémantiques. Nous proposons une analyse textométrique des verbalisations effectuées par les participants, afin de proposer des labels correspondant au mieux à leur classification.

5.5.3 Analyse des verbalisations

La textométrie s'est développée en France à partir des années 1970. Cette discipline s'appuie sur des méthodes d'analyse des données et développe de nouveaux modèles statistiques pour rendre compte de caractéristiques significatives des données textuelles [LSB94]. Nous utilisons le logiciel Txm [34] [HMP10], qui permet de faire de l'analyse statistique textuelle.

Ce logiciel a été utilisé sur les verbalisations des participants pour regrouper certains mots dans un processus de lemmatisation. Nous regroupons ainsi les verbes conjugués (*couler, coule, coulent*), ou les adjectifs accordés (rapide, rapides). Nous avons également regroupé manuellement les formes verbales et les mots d'action ayant la même racine (*écouler, écoulement*).

L'analyse des verbalisations présente au total 28868 lemmes et 562 formes lexicales différentes. Les occurrences de ces lemmes vont de 1 à 2576. Le mot *de* est celui qui présente le plus d'occurrences, comme c'est souvent le cas dans la langue française.

Si nous excluons les mots-outils comme *de, le, et*, les termes les plus utilisés sont alors *eau* (1346 occurrences), *dans* (666), *son* (635), *bruit* (488), *liquide* (376), *douche/doucher* (391) et *goutter/gouttes* présentant 234 occurrences. Ces différents mots définissent de façon générale les verbalisations effectuées par les participants à partir des sons du corpus. Nous allons utiliser les

34. `http://textometrie.ens-lyon.fr/`, TXM 0.7.2

classes issues de notre analyse afin de leur affecter un label et de produire un portrait sémantique de chacune d'elle.

Analyse lexicale des classes

Les verbalisations produites par chaque participant sont agrégées en fonction des sons. Pour chaque son, nous considérons l'ensemble des classes d'appartenance. Chaque son a été placé dans une classe par chacun des 29 participants. Nous obtenons donc 29 descriptions de classes pour chaque son. Les termes utilisés pour décrire ces différentes classes sont agrégés, ce qui permet d'obtenir une description de chaque son.

À l'issue de cette étape de description de chaque son, les termes utilisés sont regroupés en fonction des classes identifiées lors de l'analyse. Pour chaque classe, nous identifions les cooccurrences lexicales, c'est-à-dire les termes utilisés plusieurs fois.

Nous cherchons de plus à extraire des termes qui permettent de distinguer les classes entre elles. Il faut ainsi faire apparaître les spécificités des classes, et non les termes qui, s'ils présentent de nombreuses occurrences, sont communs à différentes classes (par exemple le terme *eau*). Pour les analyses suivantes, nous avons omis les termes communs à l'ensemble des classes (*eau, bruit, son, liquide*). Nous calculons pour chaque terme sa spécificité [Laf84] (ce calcul est également décrit dans la documentation du logiciel *Lexico*[35]). Cette spécificité est calculée pour un terme et une classe donnée en fonction :

- du nombre d'occurrences du terme dans la classe,
- du nombre d'occurrences du terme dans le corpus,
- de la taille de la classe,
- de la taille du corpus.

La liste des labels les plus utilisés pour décrire ces classes sont présentés en annexe H.

Analyse des 3 catégories

Nous avons effectué une analyse lexicale à partir des 3 catégories principales :

- Classe **A** : Les termes propres à cette classes sont *chasse d'eau, aspirer, évacuation, syphon, vider*. Nous proposons le label suivant : **évacuation d'eau**.
- Classe **B** : Cette classe est décrite par les termes *douche, remplir, jet, couler, fort, pluie, continu*. Il semble que plusieurs notions soient présentes dans cette verbalisation. Les termes *douche, jet, couler, fort, pluie, continu, pression* se rapporte à une notion de flux d'eau tel que décrit dans le chapitre 3. Par ailleurs cette classe est aussi composée de sons de remplissage. Cette classe sera analysée à partir de ses sous-classes.
- Classe **C** : Cette classe est décrite par les termes *goutte, essorer, mouvement, main*. Nous proposons le label suivant : **goutte/mouvement**.

Analyse lexicale des 8 catégories

Dans cette analyse, les classes **A** et **C** sont équivalentes avec les classes trouvées dans l'analyse à 3 classes. Nous allons décrire les sous-classes de la classe **B** :

- Classe B_1 : Les mots caractérisant cette classe sont les mots *robinet, couler, remplir, baignoire, lavabo*. Le mot *continu* est également présent. Nous proposons le label suivant : **ouverture de robinet**.

35. http://www.tal.univ-paris3.fr/lexico/lexico3.htm

- Classe B_2 : Dans cette classe, l'adverbe *sur*, le mot *surface* ainsi que le verbe *couler* apparaissent, ainsi que le mot *filet*. Nous proposons le label suivant : **filet sur surface**.
- Classe B_3 : La classe B_3 est clairement décrite par l'action de *verser dans* un *verre* ou un *récipient* et *remplir*. On retrouve également des noms de contenant : *bouteille, contenant, carafe*. Nous proposons le label suivant : **remplissage d'un petit récipient**.
- Classe B_4 : La classe B_4 est avant tout associée à l'action de *doucher*. Les adjectifs et mots associés sont *fort, jet, pression*. Nous proposons le label suivant : **jet**.
- Classe B_5 : Nous retrouvons des mots similaires à la classe B_4 : *doucher, jet*. Les termes *doucher* et *jet* semblent plutôt associés à l'action de *remplir*. Le mot *bassine* apparaît également. Nous proposons le label suivant : **remplissage d'un grand contenant**.
- Classe B_6 : Cette classe est décrite par les termes spécifiques tels que : *ébullition, bulle et bouillir*. Nous proposons le label suivant : **ébullition**.

À partir de l'analyse des verbalisations des regroupements en 3 et 8 catégories, nous obtenons l'arbre sémantique illustré à la figure 5.19.

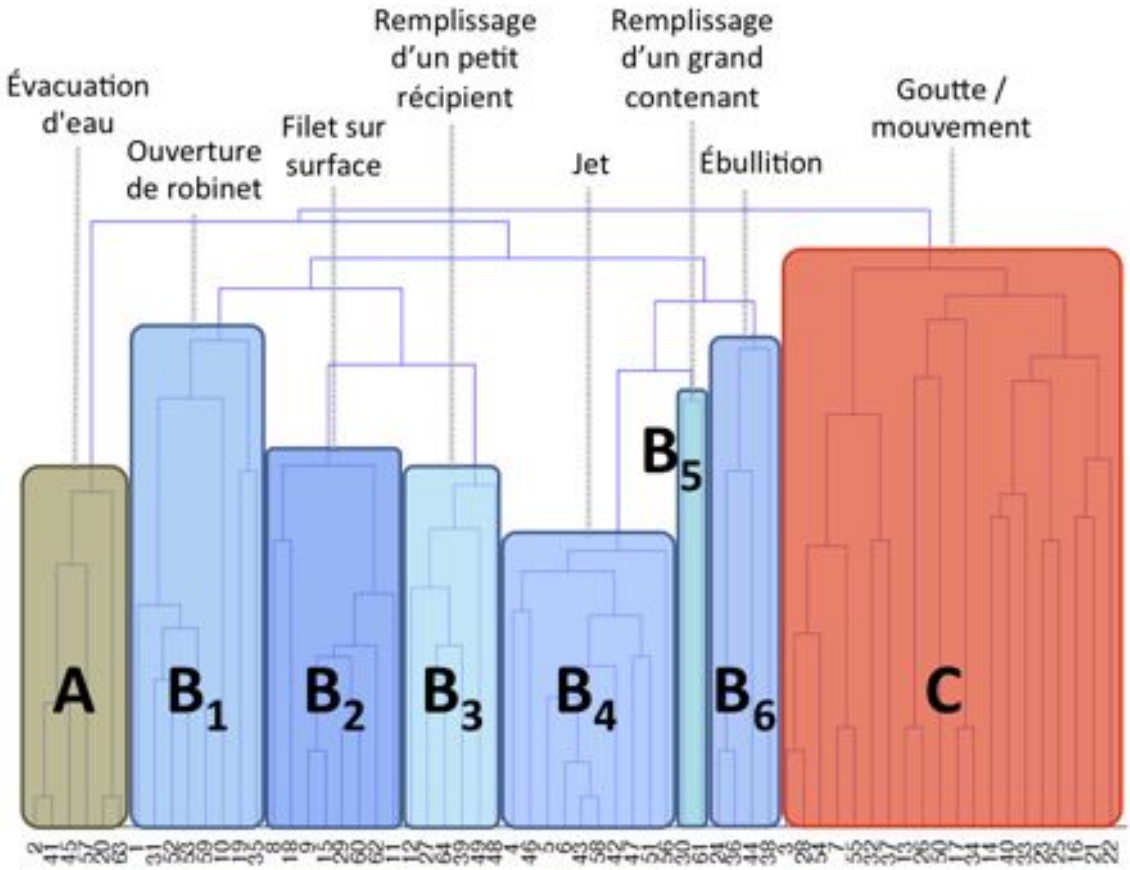


FIGURE 5.19 – Arbre des classes de sons liquides.

5.5.4 Discussion

Validité de la catégorisation

La classe **A** est une des classes les mieux définies de l'arbre. Elle est très compacte. Les sons d'eau dans des tuyaux semblent donc être caractéristiques d'un type de cause bien établi. Il est intéressant de constater que ce type de son est particulièrement lié au contexte du domicile.

La classe **B** est plus difficile à analyser. Nous pouvons nous interroger sur la validité du découpage en 6 sous-classes. Par exemple, la catégorie formée par les deux sons de la classe B_5, fusionnés à une hauteur importante, est très discutable.

La classe **C** est la plus grande des classes et contient 20 sons. Elle n'est pas très compacte et sa hauteur est très importante. Néanmoins, d'autres types de regroupements à l'intérieur de cette classe semblent difficilement envisageables.

Comparaison avec les labels des stimuli

Nous allons discuter la validité de ces catégories en comparant leurs labels avec les labels des sons les composant (présentés en annexe E) :

- La classe **A** contient 6 sons dont les labels évoquent très fortement l'évacuation d'eau : *vider un lavabo, vider un évier, tirer la chasse d'eau.*
- Au niveau des sous-classes de **B**, la classe B_1 est assez hétérogène. Elle contient des sons de robinet ouvert (*remplir un lavabo, eau coulant dans baignoire remplie*) mais aussi d'autres sons (*verser dans un bac en plastique, faire frire un oeuf, pluie tombant sur le sol depuis le toit*), respectivement les sons 10, 19, 35. La distance entre ces sons dans l'arbre est importante.
  La classe B_2 semble bien définie et contient des sons associés aux labels *remplir un gobelet en plastique, verser sur une plaque de metal.* De même, les labels des sons de la classe B_3 contiennent tous un nom de récipient *verser dans un vase vide, vider une bouteille.*
  La classe B_4 contient les labels relatifs au jet (*jet de douche, agiter un pommeau de douche*), et également certains sons de robinets ouverts (*remplir une baignoire*). Cette classe contient également un son probablement mal identifié : *steak frit sur un grill.* Dans l'arbre, nous pouvons voir que ce son numéro 43 est très fortement rapproché du numéro 58 dont le label est *jet de douche*, même si les causes provoquant ces sons sont très différentes.
  La classe B_5 ne contient que deux sons dont les labels sont *remplir un bidon en plastique* et *remplir un arrosoir.* La classe B_6 contient 4 sons dont les labels contiennent les mots *ébullition, pétillement* et *eau bouillante.*
- Les labels des sons qui composent la classe **C** contiennent le mot *goutte*, des termes de mouvement (*secouer, agiter, presser, essorer*). Certains labels sont relatifs au mouvement d'un objet dans l'eau : *objet tombant dans le bain, bouger dans le bain, mettre un glaçon dans une boisson.* Enfin deux sons évoquent plutôt la présence d'air dans l'eau : *souffler avec une paille à la surface de l'eau, bulles dans l'eau.*

Sons discrets et continus

À la différence des classes **A** et **B**, les causes des sons de la classe **C** sont en général très limitées dans le temps. Cette classe correspond principalement à des chutes de gouttes séparées, à des bulles d'air isolées ou à différents types de mouvements (secouer, mouvement dans le bain, chute d'objet).

Nous pouvons ainsi effectuer une analogie avec la perception des sons de solide [HLM+12], en considérant les classes de sons *discrets* et *continus*. Les classes **A** et **B**, contenant des sons plus continus, sont en effet regroupés dans l'arbre, bien que la hauteur de fusion soit importante.

Variation de débit d'eau

La différence entre sons discrets et continus peut s'exprimer pour certains sons en terme de débit. Les évènements sonores dont la cause physique est la chute de gouttes d'eau ou la remontée à la surface de bulles d'air produisent des sons variés en fonction du débit de gouttes d'eau ou de bulles d'air. Ces différences de débit semblent être un critère de catégorisation pertinent pour les participants. En plus des classes de sons *discrets* et *continus* que nous avons évoquées, ces variations de débit interviennent également dans les sous-classes de la classe **B**. Les classes $B2$ et B_4 se différencient ainsi par le débit du *filet* d'eau sur la surface qui devient *jet*. À l'écoute des stimuli, nous remarquons que les sons de la classe B_2, présentent un bruit de fond continu mais également des évènements discrets identifiables. Par contre, dans la classe B_4, le bruit de fond continu masque les évènements discrets difficilement audibles.

De même que le débit, la taille des élèments peut être un critère de catégorisation, comme nous l'avons vu dans la partie 5.1.3. Ainsi, en prenant en compte la taille et le débit, nous pouvons observer un certain parallélisme entre les paires de classes *filet sur surface / remplir un petit récipient* et *jet / remplir un grand contenant*.

Surface solide et liquide

Au niveau des supports sur lesquels tombe l'eau, nous avons vu que les gouttes tombant dans l'eau ou sur une surface solide sont regroupées dans la classe **C**. Par contre, dans la classe **B**, ces deux types de sons semblent avoir été séparés.

L'écoute des sons de la classe B_1, *robinet ouvert*, révèle beaucoup de sons d'eau coulant dans l'eau : les stimuli 1, 31, 52, 53, 59. La distance entre ces sons est assez petite, contrairement à la distance de fusion des trois autres sons.

Pourtant la notion d'eau coulant ou tombant dans l'eau n'apparaît pas dans l'analyse des verbalisations, contrairement à la notion de surface solide qui est très visible dans la classe B_2, ou dans la classe B_4 avec le mot *sol*. L'analyse des verbalisations individuelles peut expliquer en partie ce phénomène. D'une part les formulations de l'eau tombant dans l'eau sont variées : *de l'eau se déversant dans de l'eau, eau qui coule dans un récipient déjà rempli d'eau, dans un liquide*. D'autre part les termes utilisés pour décrire la matière dans laquelle tombe l'eau (*eau, liquide*) ne sont pas spécifiques à cette classe.

Ces deux raisons ne suffisent peut être pas complètement pour expliquer pourquoi l'eau coulant dans l'eau n'apparaît pas dans l'analyse des verbalisations. Nous pouvons supposer que la cause du son apparaît aux participants de manière plus claire lorsqu'elle résulte d'une interaction avec un solide, comme dans le cas d'une surface solide ou d'un récipient.

Synthèse

La figure 5.20 illustre une interprétation des résultats de notre étude de catégorisation. Nous avons ainsi considéré les classes de sons *discrets* et *continus*. De plus, nous avons regroupé les sons de remplissage. Au final nous pouvons retrouver le flux d'eau, décrit au chapitre 3 comme « eau qui coule », qui peut regrouper les catégories **robinet ouvert**, **jet** et **filet sur surface** et **remplissage**.

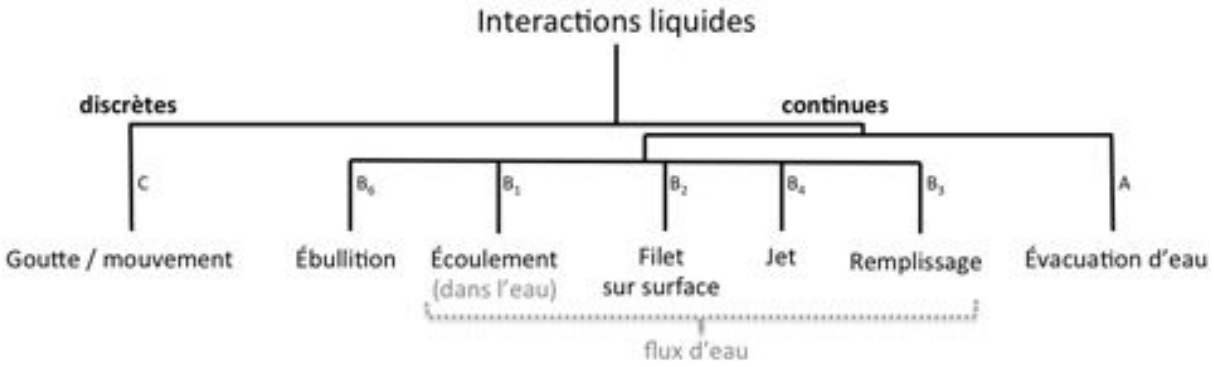


FIGURE 5.20 – Arbre des classes de sons liquides.

5.6 Conclusion

Dans ce chapitre, nous décrivons une méthodologie et les résultats d'expériences effectuées visant à obtenir des classes de sons de liquides validées perceptivement.

5.6.1 Corpus

Pour traiter l'ensemble des sons de liquide pouvant apparaître au quotidien, nous avons utilisé des outils linguistiques afin de créer un corpus contenant une variabilité de sons importante. Pour des raisons pratiques, nous avons éliminé les sons redondants. Nous avons ainsi obtenu un corpus varié et concrètement utilisable dans des expériences perceptives.

Il est toutefois assez difficile d'obtenir un corpus présentant des classes de sons équilibrées, d'autant que ces classes ne sont pas connues à l'avance. La différence de nombre de sons dans les catégories est un critère qui pourrait influencer les catégories obtenues. Il semble pourtant délicat de s'abstraire de cette contrainte. Une solution pourrait être de réitérer l'expérience de catégorisation en adaptant la taille du corpus aux réponses des participants. Nous pourrions ainsi utiliser un nombre de sons fixe pour chaque classe obtenue, et expérimenter la stabilité de ces catégories lors d'une nouvelle expérience. Nous pourrions également considérer une expérience de reconnaissance avec des catégories fixées. Cette dernière expérience de classification par choix forcé permettrait en outre de valider les labels proposés.

5.6.2 Similarités

Dans les expériences de classification, plusieurs types de catégories peuvent être utilisées, notamment les similarités acoustiques, causales, ou contextuelles. Nous avons tenté de favoriser au maximum l'utilisation de la similarité causale dans l'expérience de catégorisation. Nous avons ainsi effectué une expérience d'égalisation écologique afin de modifier la sonie des sons et de rendre le corpus plus réaliste. Nous avons également effectué une expérience d'identification afin de supprimer les sons mal identifiés.

Au final, il semble que la similarité causale ait été principalement utilisée, comme le montrent les labels des catégories obtenues. Pourtant, malgré toutes ces précautions, d'autres types de similarités sont utilisés. Par exemple, le dendrogramme montre une petite distance de fusion entre les sons 43, 58, dont la cause est très différente (friture et pluie), et qui semblent être

regroupés par un nombre important de participants selon des critères acoustiques. Il semble clair que la cause de ces sons n'a pas été identifiée correctement. L'utilisation de la similarité causale pourrait être éventuellement améliorée lors d'une expérience future par une suppression plus importante des sons mal identifiés.

De plus, les verbalisations montrent des classes plus importantes basées sur des critères acoustiques (*ce groupe est celui des bruyants*). D'autres classes sont également formées selon d'autres types de similarités (*sons « ménagers » que l'on peut retrouver dans une salle de bain*). Les participants mélangent les types de similarité au sein de leur tri (ce phénomène est appelé *cross-classification* [RM99]). Aussi les participants ayant utilisé différents types de similarités ne se démarquent pas du groupe. Le rapprochement de certaines classes de l'arbre, par exemple **Jet** avec **Ébullition** semble être aussi la conséquence de similarités acoustiques. Nous pourrions imaginer dans des expériences futures que la consigne donnée aux participants pourrait être encore plus directive, par exemple en donnant des exemples de classes à ne pas effectuer par similarité acoustique ou contextuelle.

Toutefois, il est difficile d'isoler complétement les différents types de similarité. Par exemple, les critères acoustiques des sons dépendent en général de l'action effectuée. Ces aspects causaux et acoustiques sont donc difficiles à différencier dans les catégories obtenues, même s'ils sont visibles dans les verbalisations. Par exemple, la catégorie effectuée à partir des *sons bruités* pourrait contenir exactement les même stimuli que la catégorie des *jets d'eau puissants*. De même les dissociations sons discrets/continus sont causales, mais ont des conséquences acoustiques.

5.6.3 Catégories obtenues

À l'issue de notre analyse, nous obtenons les 7 catégories de sons suivantes :
- gouttes/mouvement,
- ébullition,
- écoulement (dans l'eau),
- filet sur surface,
- jet,
- remplissage,
- évacuation d'eau.

Plusieurs de ces catégories sont associées au flux d'eau décrit dans le chapitre 3, dont les sons de remplissage. Ces derniers sont en effet très proches des catégories associées aux filets d'eau sur des surfaces ou à des jets d'eau de type douche.

Au niveau de l'analyse automatique, les catégories associées au flux d'eau sont potentiellement détectables grâce à notre système de reconnaissance décrit au chapitre 3. De même, la catégorie goutte/mouvement correspond bien aux sons issus d'un modèle physique simplifié tel que décrit dans le chapitre 4 et pourrait être reconnue par le système correspondant. Deux catégories, qui n'ont pas été identifiées dans les chapitres précédents apparaissent à l'issue de cette analyse : les catégories des sons d'ébullition et d'évacuation d'eau. Nous allons proposer dans la partie suivante, et en conclusion de cette thèse, un premier pas vers la fusion de nos méthodes pour détecter, à partir de ces différentes catégories, l'ensemble des sons de liquide du quotidien.

Épilogue

1 Vers une fusion des contributions

1.1 Introduction

Nos expériences sur la perception des sons de liquide décrites dans le chapitre précédent nous ont conduits à l'identification des catégories suivantes :

- gouttes/mouvement,
- ébullition,
- écoulement (dans l'eau),
- filet sur surface,
- jet,
- remplissage,
- évacuation d'eau.

Ces catégories se divisent en deux branches : les sons discrets et les sons continus. Nos deux approches de reconnaissance automatique semblent ainsi comparables aux stratégies cognitives utilisées par les humains lorsqu'ils reconnaissent les sons produits par l'eau. Plus précisément, plusieurs catégories correspondent à la définition du flux d'eau donné en chapitre 3 : *écoulement (dans l'eau), filet sur surface, jet, remplissage*. Ces catégories semblent donc pouvoir être reconnues par notre détecteur de flux d'eau. De même, la classe *gouttes/mouvements* coïncide avec les sons potentiellement détectables par notre système basé sur la vibration des bulles d'air décrit dans le chapitre 4. Par contre nos systèmes n'ont pas été testés sur deux autres catégories : la catégorie des *ébullitions* et celle des *évacuations d'eau*.

1.2 Validation de nos systèmes sur les catégories perceptives

Afin d'appliquer les résultats obtenus dans le domaine de la perception sur nos deux approches de détection, basées d'une part sur l'analyse du signal, d'autre part sur un modèle acoustique, nous allons conclure cette thèse par une dernière expérience. Dans cette expérience, nous proposons de tester nos deux systèmes sur une tâche de détection de sons d'eau effectuée sur le corpus de sons de liquides créé au chapitre 5.

Cette expérience permet d'utiliser nos systèmes sur un corpus contenant très variés. Le corpus de notre expérience perceptive a en effet été construit pour refléter la variété la plus importante possible des sons de liquides pouvant d'une part se produire au domicile, d'autre part être identifiée par un humain. Ce corpus contient potentiellement des sons ne se trouvant pas dans nos précédentes expériences. De plus les catégories de sons de liquides ont été identifiées sur ce corpus. Cette expérience nous permet donc d'obtenir directement des résultats de reconnaissance par catégorie.

Par ailleurs, il faut préciser que les sources sonores qui composent ce corpus sont présentées

de manière isolée, contrairement à nos expériences précédentes. Néanmoins, la robustesse de nos systèmes à détecter les sons de liquides dans des scènes sonores de la vie réelle a déjà été éprouvée dans les expériences des chapitres 3 et 4.

1.3 Mise en œuvre

Nous utilisons comme détecteur de flux d'eau le système initial décrit au chapitre 3. Nous avons supprimé l'étape de post-traitement de ce système, qui supprime les sons d'une durée inférieure à trois secondes.

Le système de détection de bulles d'air que nous utilisons a été décrit au chapitre 4. Pour que ce système puisse détecter des bulles d'air isolées, (créées par exemple par la chute d'une goutte d'eau unique), l'étape de post-traitement qui vise à supprimer les bulles d'air isolées n'est pas effectuée.

Nous appliquons ces deux systèmes sur le corpus de l'expérience perceptive, qui contient 64 évènements sonores d'une durée moyenne de 3 secondes. Les évènements sonores sont concaténés dans un fichier audio et espacés par un silence d'une seconde. Un évènement sonore est considéré reconnu par l'une des deux approches si le système correspondant détecte un son d'eau pendant la durée de l'évènement sonore.

La figure 2 présente les résultats de cette expérience. Les bandes de couleur signifient respectivement que le son a été détecté par le détecteur de flux d'eau (en bleu), le détecteur de bulles d'air (en rouge), les deux détecteurs (en vert), ou aucun des deux (en jaune).

2 Résultats

2.1 Résultats globaux

Dans cette expérience effectuée sur les 64 sons de notre corpus de sons de liquide du quotidien, nous obtenons les résultats suivants :

- 21 sons détectés uniquement comme flux d'eau,
- 15 sons détectés uniquement comme bulle d'air,
- 20 sons détectés par les deux approches (flux d'eau et bulle d'air),
- 8 sons manqués.

Nous avons donc au total une détection correcte de 87% des sons du corpus. Compte tenu de la variabilité acoustique de ces évènements sonores, ce score global semble satisfaisant.

2.2 Résultats par catégorie

Les sons de la classe A (*évacuations d'eau*) ont été majoritairement détectés par nos deux détecteurs. Ceci peut s'expliquer car les bruits d'évacuation d'eau présentent à la fois les caractéristiques du flux d'eau (son continu assez bruité) et également la présence de gouttes. Ce phénomène est particulièrement audible pour les sons de chasse d'eau (20 et 63). Par contre, un son n'a pas été détecté (le numéro 45, dont le label est *vider un évier*). À l'écoute il semble que ce son, plutôt grave, soit produit par un siphon ouvert alors que l'évier est complétement rempli. Notre système de détection de flux d'eau, basé sur la couverture spectrale qui privilégie les fréquences aiguës, ne permet pas de détecter ce son caractérisé par des basses fréquences.

La classe B_1 (*ouverture de robinet*) a été entièrement détectée comme flux d'eau, et à moitié comme bulle d'air. Nous avons mentionné au chapitre 5 que cette classe est plutôt caractérisée par de l'écoulement de l'eau dans l'eau. Ce phénomène physique implique une présence importante

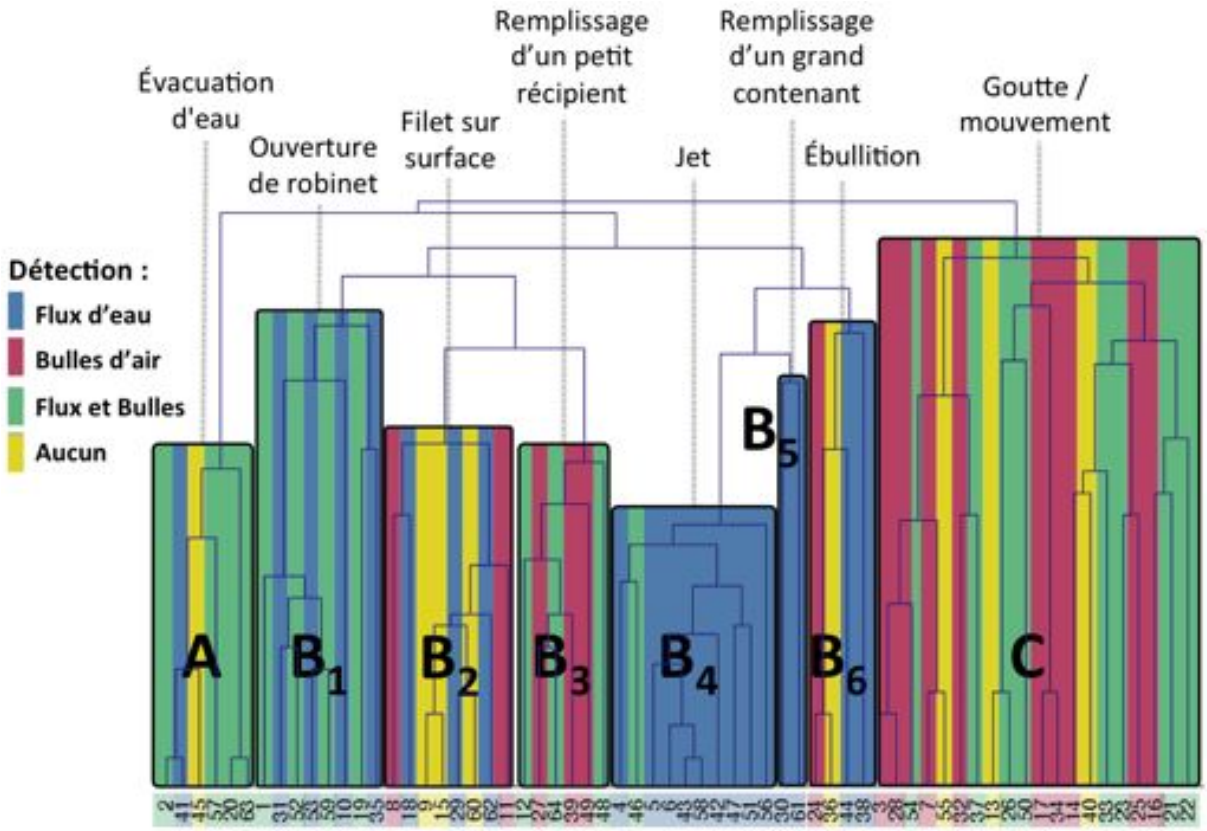


FIGURE 2 – Résultats de nos deux détections sur les catégories perceptives.

de bulles d'air. Pourtant, certains sons, présentant des écoulements plus importants, ne sont pas reconnus par notre détecteur de bulles. Les sons de cette catégorie permettent donc d'identifier une limite concrète ente les sons présentant des bulles d'air identifiables par notre approche, et les sons d'eau dont les sons de bulles d'air ne sont pas assez saillants.

La classe B_2 (*filet sur surface*) est la classe la moins bien détectée, proportionnellement à son nombre d'éléments. Les trois sons manqués sont définis par les labels suivants : *verser de la peinture dans un pôt* (9), *verser sur une plaque de métal* (15), *filet d'eau dans un évier* (60). Les filets d'eau de certains sons semblent présenter une intensité trop faible pour être détectés comme des flux d'eau. De plus la surface métallique sur laquelle tombe ce filet produit un bruit important qui masque les éventuels sons de bulle. Enfin, la peinture, considérée comme un liquide très visqueux, ne semble pas provoquer de bulles d'air.

La classe B_3 (*remplissage d'un petit récipient*), a été entièrement détectée comme bulle d'air, et à moitié comme flux d'eau. À l'écoute, ces sons de remplissage présentent une forte présence de bulles d'air. Au delà de la montée en fréquence produite par le son de remplissage, cette catégorie est caractérisée par le filet d'eau se déversant dans l'eau qui remplit déjà le petit récipient. Cette classe est donc comparable à la classe B_1, mais le débit d'eau est ici moins important. Ainsi, contrairement à la classe B_1, le système de détection de bulle d'air prend ici le dessus sur le système de détection de flux.

Les classes B_4 et B_5 présentent des jets d'eau bruités, trop complexes pour notre détecteur

de bulles d'air. Par contre, les sons qui les composent ont tous été détectés comme flux d'eau.

Dans la classe B_6 (*ébullitions*), trois sons ont été détectés et un son a été manqué. L'écoute de ces stimuli permet de comprendre le résultat des détections. Une fois encore la granularité des évènements sonores semble avoir un impact très important sur l'acoustique des sons et les résultats de la détection. Dans les sons d'ébullition, les évènements sonores sont des bulles d'air remontant à la surface. L'un des sons est caractérisé par des bulles de taille importante et plutôt isolées : *ébullition dans une grande casserole* (24). D'autres sons résultent d'un ensemble de bulles important qui produisent un son bruité et continu : *eau bouillante dans une casserole* (44), et *pétillement après ouverture d'une cannette* (38). Un autre son, *ébullition dans un chaudron* (36), présente un nombre élevé de bulles de taille importante et constitue un intermédiaire entre ces phénomènes, qui n'est ici détecté par aucune de ces deux méthodes.

La classe C est la classe des gouttes et des mouvements. Parmi ces 20 sons, 17 ont été détectés comme sons présentant des bulles d'air. Les sons non détectés sont caractérisés par une vibration importante d'un élément solide : *gouttes dans un évier en métal* (55), *secouer une bouteille en plastique remplie* (13), *agiter un récipient en métal à moitié rempli* (40).

2.3 Classes de sons discrets et continus

Nous observons une tendance globale pour les classes de sons continus (les catégories A et B) à être détectées comme flux, et pour la classe discrète (catégorie C) à être détectée par notre détecteur de bulles. Ainsi, 72% des sons de la classe des sons continus ont été détectés comme flux, et 85% de la classe discrète a été identifiée comme bulle.

Ce phénomène est pourtant moins visible que nous aurions pu l'imaginer. Nous pouvons ainsi observer de nombreux sons détectés par les deux approches. De plus, la catégorie *remplir un petit récipient*, considérée comme proche du flux d'eau, est au final mieux reconnue par notre détecteur de bulles.

Les résultats de nos détecteurs permettent donc d'observer un autre type de classement des sons, qui semble plutôt dépendant de la granularité des bulles d'air et diffère sensiblement des catégories perceptives basées sur la cause physique des sons.

3 Discussion

Au final, la majorité des catégories ont été reconnues de manière satisfaisante. Les deux classes qui n'ont pas été testées lors de nos précédentes expériences, les classes de sons d'évacuations et d'ébullitions, ont été reconnus à plus de 75%.

La classe des ébullitions semble similaire par la variété à celles des sons produits par un écoulement. Elle est constituée de sons potentiellement détectables par l'un des deux systèmes, selon le nombre et la taille des bulles d'air remontant à la surface. Un son manqué dans les résultats illustre néanmoins un manque sur l'une de ces deux approches. La classe des évacuations est reconnue en partie par notre détecteur de flux. Toutefois, un son très sourd n'a pas été reconnu, ce qui montre une limite à cette méthode, car le descripteur utilisé est plutôt sensible aux fréquences aiguës produites par le flux d'eau. Parmi les sons non reconnus dans les autres classes, certains font intervenir une vibration importante de matériel. D'autres sont produits par un liquide très visqueux qui ne produit pas de bulle d'air.

Ces sons manqués constituent la limite actuelle de nos systèmes. S'ils peuvent être potentiellement identifiés par des humains, il sont néanmoins complexes et plutôt rares. Nous pouvons supposer que ces sons manqués ne sont pas déterminants pour une application de reconnaissance d'activités de la vie quotidienne.

Ainsi, malgré quelques erreurs, la reconnaissance des sons d'eau à partir de deux approches, l'une basée sur des sons continus et l'autre sur des sons localisés, semble-t-elle très pertinente. D'une part, les 87% d'identifications correctes des évènements sonores montrent l'efficacité et le caractère suffisant de nos deux systèmes. D'autre part, plus de la moitié des sons n'ont été identifiés que par une seule des deux méthodes. Ce résultat confirme l'aspect nécessaire de ces deux types d'approches différentes.

Conclusion

1 Cheminement de recherche

1.1 Reconnaissance et caractérisation des sons d'eau

Nous avons étudié dans cette thèse les phénomènes sonores produits par l'eau. Ce thème de recherche a été initialement guidé par le projet ANR IMMED, dont l'objectif est la reconnaissance d'activités pour l'aide au diagnostic. Les sons liés à l'utilisation de l'eau permettent d'inférer sur différentes activités du quotidien effectuées au domicile, citées par les médecins comme activités d'intérêt. Les sons d'eau apparaissent notamment dans les activités liées à l'alimentation, à l'entretien du domicile ou à l'hygiène.

Le projet IMMED nous a donc amenés à travailler sur la détection automatique de sons produits par l'eau dans des fichiers audio. Cette tâche peut être placée dans le contexte plus global de la reconnaissance automatique de sources sonores. Nous avons vu dans le chapitre 2 qu'un domaine récent, appelé analyse computationnelle de scènes sonores, propose un cadre pour les recherches menées sur les enregistrements de la vie réelle. Ce type d'enregistrement présente des caractéristiques différentes des corpus audio produits en studio, notamment par la présence de sons environnementaux et une superposition fréquente des sources sonores. Des applications récentes proposent, à partir d'enregistrements effectués dans la vie quotidienne, la détection des sons environnementaux.

Nous avons parcouru au chapitre 3 les travaux de recherche sur la détection de sons d'eau. Ceux-ci semblent néanmoins difficilement adaptables à la problématique du projet IMMED où les lieux d'enregistrements sont très variés. À partir d'une analyse « bas niveau » orienté sur le choix de descripteurs audio, nous avons proposé un nouveau descripteur, sensible au son de flux d'eau dans un environnement bruité. Ce descripteur est utilisé dans un système assez simple basé sur des seuils, et permet de détecter les sons de flux d'eau de façon plus robuste que les méthodes classiques. Ce système a été intégré dans le projet IMMED, et contribue à la reconnaissance d'activités de la vie quotidienne. Par ailleurs, une amélioration des performances du système est obtenue par l'ajout d'une étape de classification.

Malgré cette dernière amélioration, notre système de détection de flux d'eau n'est pas adapté à la reconnaissance de l'ensemble des sons d'eau produits par les activités du domicile. Certaines activités provoquent en effet d'autres sons que le flux d'eau, par exemple des gouttes ou des mouvements d'eau. Ces évènements sonores sont pourtant facilement identifiables par un humain comme résultant d'une interaction avec un liquide. Pour comprendre la particularité de ces sons, nous nous sommes rapprochés des études d'acoustique qui théorisent les phénomènes vibratoires intervenant dans la production de sons de liquide. Le phénomène physique à l'origine des sons de

liquide est la vibration des bulles d'air dans l'eau. Différentes études de synthèse sonore proposent notamment de créer un ensemble de sons de liquide diversifié à partir du calcul des vibrations de bulles d'air.

Nous avons proposé dans le chapitre 4 un système de détection des sons produits par l'eau fondé sur la reconnaissance acoustique des vibrations de bulles d'air. Ce système nous permet de détecter des sons d'eau assez variés, dont les sons produits par des activités domestiques. Pourtant, si au niveau de l'acoustique les vibrations de bulles d'air permettent d'unifier les sons de liquide autour d'une origine commune, notre système basé sur la détection de bulles d'air ne peut détecter l'ensemble des ces sons. En effet, les sons produits par une grande quantité de bulles d'air, typiquement les flux d'eau, semblent trop complexes pour être reconnus par cette approche.

Au final les systèmes de détection de flux d'eau et de bulles d'air sont complémentaires, car ils se trouvent respectivement limités par des évènements trop localisés dans le plan temps/fréquence ou des sons trop bruités. Pour comparer notre approche, basée sur ces deux systèmes, avec le fonctionnement humain, nous avons effectué une expérience perceptive. Cette expérience vise à identifier les catégories utilisées par les humains dans leur représentation du monde sonore. En effet, les personnes qui écoutent une scène sonore jugent si les sons qui la composent peuvent être une instance d'une catégorie connue. La recherche des catégories permet donc de mettre en évidence différentes stratégies cognitives visant à identifier les sons.

Nos expériences, basées sur la similarité causale entre les sons, nous a conduits à l'identification des catégories des sons de liquides suivantes : *gouttes/mouvement, ébullition, écoulement (dans l'eau), filet sur surface, jet, remplissage, évacuation d'eau*. Ces catégories peuvent être décomposées en deux classes, celle des sons continus et celle des sons discrets. Nous pouvons donc, dans une certaine mesure, identifier deux types de stratégies cognitives utilisées par les humains lorsqu'ils reconnaissent les sons produits par l'eau, qui sont comparables à nos deux approches de détection automatique.

Les catégories trouvées permettent de tester nos systèmes sur un nouveau corpus de sons de liquide. Ce corpus a l'avantage d'être organisé en classes de sons et de refléter la variété des évènements sonores du quotidien. Une expérience utilisant nos deux systèmes sur ce corpus atteste de l'aspect nécessaire et suffisant des deux types d'approches de détection. Toutefois, les résultats de nos systèmes ne correspondent pas exactement aux deux classes de sons discrets et continus. Nous obtenons au final trois types de caractérisation des sons d'eau, issus :

- des modèles acoustiques, dans lesquels les sons d'eau sont principalement décrits comme un ensemble de bulles,
- des catégories perceptives de sons de liquides, basées sur la cause physique du son en terme d'action,
- de nos deux approches de détection à partir du signal sonore, qui mettent en évidence différentes granularités de bulles d'air dans un ensemble allant de la vibration de la chute d'une goutte d'eau unique à un flux bruité et continu.

Ainsi, les différentes thématiques de recherche abordées ont permis d'améliorer la caractérisation des sons d'eau, et donc les stratégies de reconnaissance.

1.2 Le triptyque de la recherche sur les phénomènes sonores

Ces résultats se sont appuyés sur une lecture transversale d'études issues de différents domaines. Nous avons cité plusieurs travaux de reconnaissance automatique de sons d'eau, des

travaux d'acoustique avec leurs applications à la synthèse sonore, et des travaux de perception des sons de liquide dans le contexte des sons environnementaux. Pourtant, à notre connaissance, les passerelles entre ces différentes thématiques n'avaient jusqu'alors pas été mentionnées : ces différentes études ne semblent pas citées en dehors de leur domaine de recherche. Cette thèse constitue donc le premier travail sur les sons d'eau qui unifie ces champs de recherches.

Les sons d'eau ont ainsi servi de fil conducteur lors du parcours de trois domaines d'étude du phénomène sonore : le signal, l'acoustique et la perception. Le développement des recherches dans chacune de ces thématiques a découlé des problématiques rencontrées.

Ces différents domaines permettent d'analyser la production du son, les caractéristiques du signal sonore, et la manière dont il est interprété par un auditeur. Ils constituent un triptyque qui semble adéquat à l'obtention d'une vision globale d'un phénomène sonore. Le domaine informatique dans lequel s'inscrit cette thèse a servi de support à l'utilisation et au développement de divers outils, dont les méthodes automatiques appliquées au signal sonore, les interfaces de passation des expériences perceptives, et les programmes d'analyse des résultats. Au final, ces différentes thématiques se sont mutuellement influencées, et ont engendré différentes contributions. Si nous nous sommes concentrés au cours de cette thèse sur les évènements sonores produits au sein du domicile, nous espérons que les analyses et les méthodes développées pourront être appliquées aux sons d'eau de la nature, qui marquent fortement un paysage sonore et peuvent provoquer chez l'auditeur un panel d'émotions variées.

2 Applications

2.1 Reconnaissance des sons d'eau au domicile

Application à l'identification d'activités

Nos systèmes de reconnaissance automatique de sons d'eau produisent des résultats satisfaisant dans un contexte de vie quotidienne. Bien que les résultats de nos précédentes expériences dévoilent certaines erreurs, ces dernières apparaissent surtout en présence de sons très spécifiques qui interviennent rarement dans les activités d'intérêt. Nos systèmes de reconnaissance peuvent donc être appliqués à l'identification des activités du quotidien.

Nous avons ainsi cité au chapitre 3 le cadre applicatif du suivi à distance des activités d'hygiène d'une personne âgée atteinte de trouble du comportement [CKZ+05]. La détection de sons d'eau par un microphone évite des situations embarrassantes d'observation directe ou d'utilisation de la vidéo. Elle peut contribuer chez les personnes atteintes de démence, par exemple dans le cadre du système COACH [MBCH08], à l'obtention d'un résumé des activités effectuées dans la journée [TSGM10]. Elle est également proposée dans le cadre du maintien à domicile des personnes isolées [FAH06].

Notre détecteur de flux d'eau produit des résultats tout à fait utilisables pour la détection robuste de sons d'eau dans le cadre de la reconnaissance d'activité. Il peut être appliqué directement à différents types de domicile. Une étape de classification permet également d'améliorer les résultats, mais nécessite un apprentissage préalable des modèles. Notre système de détection de bulle d'air est par ailleurs déterminant pour la reconnaissance de certaines activités : *vaisselle à la main, bain.*

Si le projet IMMED a illustré la pertinence de notre approche dans le cas d'un dispositif portable, les applications décrites ci-dessus utilisent un microphone fixe. Ce contexte devrait encore améliorer les résultats de nos méthodes de reconnaissance. D'une part la position du

microphone pourra être mieux adaptée aux sons à reconnaître, et éloignée des sources potentielles de bruits parasites. D'autre part, la localisation du microphone dans la cuisine ou la salle de bain privilégie la reconnaissance d'activités d'intérêt, au détriment d'autres activités pouvant provoquer des fausses alarmes.

Enfin, bien que cette tâche ne soit pas l'objet de cette thèse, la reconnaissance de plusieurs débits d'eau différents est évoquée dans ces perspectives, et semble envisageable à partir de nos approches. La reconnaissance du débit pourrait ainsi être appliquée à l'amélioration de la reconnaissance d'activités, mais aussi au suivi de la consommation d'eau.

Application à la prévention du gaspillage d'eau

Nous avons ainsi cité d'autres applications de la reconnaissance de sons d'eau au domicile, cette fois plutôt orientées sur des objectifs écologiques. Pour ce type d'applications, un autre avantage de nos systèmes réside dans leur faible coût en terme de temps de calcul. Nos deux systèmes se basent en effet sur une approche bas niveau, que ce soit au niveau de notre descripteur, la couverture spectrale, ou de notre analyse du spectrogramme. Nos approches sont donc utilisables en temps réels.

L'utilisation d'un microphone près du robinet, associée à la tâche de détection de son d'eau, permet de prévenir les gaspillages. Nos deux approches permettent de détecter des consommations d'eau variées, allant du goutte à goutte aux flux importants. Elles pourraient tout à fait être utilisées dans ce cadre de prévention du gaspillage d'eau *inter-activité* et *intra-activité* tel que décrit dans [VSN+11].

Au delà de l'utilisation du robinet, une autre étude montre que l'affichage de consommation d'eau pendant une douche permet aux utilisateurs de modérer leur utilisation [KG09]. Nos systèmes, adaptés à la reconnaissance de différents débits, pourraient également être utilisés pour afficher la quantité d'eau consommée en temps réel.

Enfin, par rapport à un système basé sur un compteur d'eau tel que présenté dans [KG09], la reconnaissance de l'utilisation de l'eau à partir du son permet, outre une installation plus facile, le suivi de plusieurs arrivées d'eau à partir d'un microphone unique. Un détecteur utilisant un microphone unique pourrait donc être utilisé pour des sanitaires collectifs ou des usines, et pourrait permettre, à distance, de suivre la consommation d'eau ou de liquide. Des détecteurs installés dans les bâtiments non occupés pourraient également signaler une fuite.

Dispositif portable

Le projet IMMED nous a conduits au développement de méthodes et d'algorithmes adaptés aux environnements sonores produits par les dispositifs portables. Nos systèmes, qui peuvent être utilisés en temps réel, peuvent tout à fait être intégrés dans un dispositif portable, par exemple sous la forme d'une application pour smartphone. L'avantage d'un tel dispositif est qu'il ne nécessite aucune installation matérielle.

Ce logiciel pourrait d'une part présenter un aspect ludique : la comparaison entre notre perception du monde sonre et la détection de sons d'eau. Il pourrait d'autre part être utilisé pour des besoins précis : la surveillance d'une fuite d'eau à distance ou l'émission d'un signal sonore après obtention de l'ébullition dans un contexte de cuisine.

2.2 Reconnaissance des sons d'eau à l'extérieur

Il semble également tout à fait envisageable d'utiliser nos outils de reconnaissance de sons d'eau à l'extérieur. Des dispositifs commerciaux proposent par exemple la détection de chute

dans un pisicine[36], à partir de la détection de mouvement dans l'eau ou de l'immersion d'un capteur portable. Ces dispositifs pourraient être améliorés ou remplacés par une détection des sons d'eau à partir d'un microphone.

Nous avons vu que l'eau produisait une variété importante de sons dans la nature. Nos approches semblent pouvoir être appliquées à ces sons de liquide variés. Dans ces environnements extérieurs, certains évènements sonores pourraient provoquer des fausses alarmes et nécessiter éventuellement une modification des paramètres de nos systèmes. Néanmoins, notre approche basée sur un modèle de son continu et sur un modèle de sons produits par les bulles d'air semble convenir pour reconnaître d'autres classes de sons d'eau, comme les sons de ruisseau, de rivière, de cascade, de pluie, mais aussi les mouvements d'eau provoqués par la mer, ou le passage d'un bateau. Ce champ de recherche ouvre d'autres perspectives applicatives.

Comme nous l'avons cité dans le chapitre 2, la détection de son a été exploitée en écologie, notamment avec la détection du passage de poissons migrateurs qui ont la particularité de donner des battements de queue sonore [DNM+13]. Le comptage des individus permet de suivre l'évolution de l'espèce. Nos systèmes semblent utilisables pour ce type d'application, mais peuvent nécessiter une adaptation pour rester robustes au bruits d'écoulement de la rivière.

Les techniques de reconnaissance de sons d'eau peuvent également être utilisées pour l'annotation de collections de paysages sonores. Le *World Soundscape projet*, décrit dans le chapitre 2, a pour objectif la conservation du patrimoine sonore. Pour ce type de corpus, les outils d'annotations automatiques sont très utiles. Une reconnaissance des sons d'eau peut ainsi permettre d'annoter les sons provenant de la mer, de la pluie, ou d'eau vive.

Des enregistrements de ce type pourraient également avoir un intérêt pour la documentation et la recherche dans le cadre du climat et de la météo. Des annotations automatiques sont par exemple effectuées dans le projet Eurequa[37]. D'autres projets de recherches actuels utilisent également des enregistrements extérieurs de longue durée, notamment en milieu urbain, qui nécessitent une annotation manuelle ou semi/automatique : *Ciess*[38], *SensorCity*[39], *SenseCity*[40]. Pour ces derniers projets, l'annotation automatique de la pluie, ou des sons d'eau provoqués par exemple par une chaussée mouillée, peut constituer une application de nos systèmes.

3 Perspectives de recherche

Nous présentons dans cette partie les perspectives de recherche et de développement de nos contributions. Nous avons ainsi considéré une analyse acoustique de notre corpus de sons de liquide, la fusion de nos deux approches de détection, l'identification directe des activités du quotidien à partir des sons d'eau, ainsi que des perspectives de recherches théoriques pour chacun des thèmes de recherche abordé.

3.1 Analyse acoustique

L'expérience décrite dans l'épilogue constitue une étape vers l'analyse acoustique des sons d'eau. Nous avons ainsi remarqué que nos deux approches de détection permettent d'identifier des classes de sons d'eau, dont le découpage diffère légèrement des catégories établies perceptivement.

36. Par exemple sur le site http://www.aquasensor.com/
37. http://eurequa.univ-tlse2.fr
38. http://www.petra.univ-tlse2.fr/ciess
39. http://www.sensorcity.nl/
40. http://sense-city.univ-paris-est.fr/

Nous supposons que ces classes sont formées en fonction de la granularité de la vibration des bulles d'air ou des impacts d'eau sur des surfaces solides.

Un prolongement de ce travail pourrait être la projection de ces sons dans un espace acoustique adapté, basé sur des descripteurs bas niveau, qui permettrait de faire apparaître ces classes. Ce travail de projection des catégories perceptives dans un espace acoustique a été effectué dans [GKW07]. Dans cette étude composée de sons très différents, incluant par exemple des vocalises, les auteurs identifient des corrélations entre certains descripteurs acoustiques et les axes issus d'un positionnement multidimensionnel sur les catégories perceptives.

Dans notre étude, caractérisée par un corpus de sons produits par le même élément, il serait intéressant de voir si cet espace existe, et quels sont les descripteurs corrélés. Il serait notamment intéressant de trouver des descripteurs capables d'exprimer la granularité des sons. La moyenne du flux spectral pour chaque son, pourrait peut être, par exemple, exprimer cette granularité.

3.2 Fusion des approches de détection

Pour certaines applications, l'annotation de segments en sons d'eau peut constituer un résultat suffisant. Nous considérons la fusion de nos deux approches pour unifier nos outils et effectuer une segmentation unique en son d'eau à partir d'un fichier audio. Nous décrivons deux types de fusion : des fusions de type tardive et une fusion intermédiaire.

Fusions tardives

Dans la fusion tardive, nous utilisons directement la sortie de nos deux systèmes, sans remettre en cause le fonctionnement de chacune des deux méthodes.

Intersection : L'intersection est un type de fusion intéressante dans le cas où les deux systèmes peuvent détecter un nombre important de sons. Cette fusion peut notamment être très utile pour supprimer des fausses alarmes. Pourtant, l'expérience présentée en épilogue nous a révélé que de nombreux sons ne sont détectés que par l'une des deux approches. Ainsi, utiliser l'intersection des sorties de nos deux systèmes n'est pas une solution intéressante pour détecter l'ensemble des sons d'eau.

Cette fusion peut néanmoins permettre de détecter des types de sons particuliers, qui ont été globalement reconnus par les deux approches. Elle pourra par exemple être utilisée pour reconnaître les sons d'écoulement d'eau dans l'eau, qui correspondent aux classes B_1 et B_3, et permettre la reconnaissance des sons produits par le remplissage d'une baignoire ou par un ruisseau.

Union : L'union des sorties des deux systèmes de détection semble être une solution intéressante pour être utilisée dans des applications de détection de sons d'eau variés. Utilisée dans l'expérience présentée en épilogue, elle permet de reconnaître une grande majorité de sons d'eau.

Néanmoins, si notre système de détection de bulle d'air a été testé de manière satisfaisante sur l'un des enregistrements du projet IMMED, la généralisation de son utilisation à l'ensemble du corpus produit des fausses alarmes. Ce système nécessiterait donc une amélioration afin de pouvoir être utilisé de manière robuste dans des conditions très variées.

Autres fusions tardives : Nous pouvons par ailleurs considérer d'autres types de fusion, qui correspondent plus spécifiquement à notre problématique. Nous pouvons ainsi supposer que des gouttes d'eau isolées apparaissent à certains instants comme l'ouverture et à la fermeture du

robinet. De même, certaines activités, par exemple le bain ou la vaisselle, produisent des sons de mouvements d'eau dans des instants relativement proches de l'ouverture du robinet et du flux d'eau.

De manière générale, nous pouvons considérer une union quand les sorties des systèmes de flux et de bulles se trouvent à l'intérieur d'une même fenêtre glissante de durée Δ_t qui pourrait être de l'ordre de la minute. La même technique pourrait également être appliquée uniquement à partir de la sortie du flux d'eau et permettrait de détecter les zones de gouttes dans un intervalle de confiance autour du flux d'eau.

Fusions intermédiaires

La fusion intermédiaire consiste à mélanger nos deux approches pour créer un système unique. Cette fusion semble nécessiter un travail plus complexe, car elle demande une modification de chacun des deux systèmes.

Nous avons vu que les sons d'eau sont en grande partie générés par les bulles d'air, et que la granularité de ces bulles produit un ensemble de sons allant d'évènements isolés à un flux continu et bruité. Tous les sons intermédiaires existent entre ces deux extrêmes. Ainsi, un système de fusion intermédiaire pourrait utiliser :

- des valeurs de couverture spectrale élevées pour reconnaître des évènements de type jet,
- des valeurs de couverture spectrale importantes ainsi que la présence de bulles d'air, pour les sons de type écoulement,
- des valeurs de couverture spectrale faibles ainsi que la présence de sons de bulle saillants pour détecter les sons d'eau de débit moins important.

Un tel système peut permettre la détection d'un ensemble des sons d'eau sans avoir à considérer nos deux approches de manière individuelle. De plus ce système pourrait permettre l'identification de trois type de sons d'eau différents, qui semblent liés à la quantité d'eau en mouvement. Dans la continuité d'une telle approche, des seuils supplémentaires pourraient être utilisés pour caractériser différents types de débit d'eau.

Conclusion

Les différents types de fusion évoqués peuvent permettre d'obtenir une segmentation unique en son d'eau à partir d'enregistrements de la vie quotidienne. Les premiers tests de fusion tardive effectués sur le corpus IMMED, dans lequel les enregistrements sont très bruités, n'ont pas donné de résultats satisfaisants, car notre système de détection de bulles d'air produit trop de fausses alarmes.

La réalisation concrète de la fusion de nos systèmes est donc considérée comme une perspective de cette thèse. La fusion intermédiaire est un objectif intéressant qui peut permettre de détecter l'ensemble des sons d'eau sans avoir à considérer les deux approches de manière individuelle.

3.3 Identification des activités

Pour d'autres applications, il peut être intéressant de conserver l'information des deux systèmes afin de caractériser le son d'eau. L'utilisation des différentes classes de sons d'eau peut apporter des indices supplémentaires pour reconnaître l'activité effectuée. Par exemple, les activités de vaisselle peuvent impliquer des mouvements d'eau, et les activités de ménages peuvent provoquer des gouttes isolées lors de l'essorage. La reconnaissance précise d'un son d'ébullition peut permettre d'inférer sur la préparation d'un plat ou d'une boisson.

D'ailleurs, nous avons travaillé au sein du projet IMMED sur l'identification directe des activités à partir des sons d'eau. Nous avons effectué un test préliminaire avec des participants, pour voir si les activités effectuées par les patients étaient potentiellement reconnaissables à partir de l'audio.

Protocole de test

Dans un premier temps, les participants doivent identifier l'activité à partir d'un extrait audio. Ils peuvent ainsi choisir l'une des réponses suivantes :

- se laver les mains,
- se laver les dents,
- préparer à manger/boire,
- jardiner,
- je ne sais pas.

Dans un deuxième temps, les participants indiquent les raisons qui ont permis de choisir cette activité. Les participants peuvent ainsi choisir des causes dans la liste suivante (établie arbitrairement, ce test ayant été effectué avant l'établissement des catégories perceptives) :

- Gouttes
- Jet / flots,
- Choc / impacts,
- Frottement,
- Versement,
- Écoulement,
- Éclaboussure / splash / plouf,
- Remplir,
- Autres.

Les participants ont également la possibilité de décrire librement d'autres raisons.

Dans un troisième temps les participants ont accès à la vidéo de l'extrait, et doivent à nouveau identifier l'activité dans la liste présentée dans la première partie.

Résultats

Dans ce test, 19 extraits d'activités issus du corpus IMMED ont été présentés. Onze participants de notre équipe de recherche ont effectué cette expérience.

Au final, si certaines activités peuvent être très bien reconnues par l'audio, comme par exemple *se laver les dents*, d'autres activités, comme *préparer à manger* sont plus difficilement identifiables. En moyenne, la similarité entre l'identification de l'activité par l'audio seul, et de son identification par l'audio et la vidéo est de 63%.

Ce score traduit des résultats d'identification assez faibles. De plus, dans ce test, les participants n'utilisent pas uniquement les sons d'eau, mais aussi les chocs (produits par exemple par la vaisselle) et la parole. Près de 10% des réponses ont ainsi été influencées par la compréhension de la parole.

Si nous considérons que le participant humain possède des capacités d'écoute et de compréhension qui le placent loin devant la machine, les résultats potentiellement atteignables par cette tâche au niveau automatique sont trop faibles pour envisager son utilisation en l'état. Par contre, nous pourrions utiliser, comme c'est le cas dans le projet IMMED, une fusion de paramètre audio

et vidéo. Les résultats de nos deux approches pourraient donc être utilisés comme paramètres d'un modèle HMM, et contribuerait à l'identification de différents types d'activités.

3.4 Perspectives théoriques pour chaque axe de recherche

Dans cette thèse nous avons étudié un phénomène sonore de manière transversale à plusieurs axes de recherche. Nous détaillons dans cette partie les contributions des méthodes développées et les perspectives de recherches théoriques pour chacun de ces axes.

Approche signal

Au niveau de la détection de flux d'eau, détaillée au chapitre 3, nous avons travaillé sur la reconnaissance de sons environnementaux pour la reconnaissance d'activités dans un corpus de vidéo bruité. Notre approche « bas niveau » basée sur le choix de descripteurs pertinents pour modéliser ce phénomène nous a amenés à l'élaboration d'un nouveau descripteur, la couverture spectrale.

La généralisation de l'utilisation de ce nouveau descripteur à d'autres applications semble envisageable. Nous avons vu, par exemple qu'il produisait des résultats satisfaisants pour l'aspirateur. D'autres évènements sonores, comme les sonneries ou le chant des oiseaux, sont caractérisés par des valeurs de couverture spectrale importantes.

De plus, l'approche développée a rapidement permis d'obtenir des résultats bien supérieurs aux méthodes classiques. Nous pouvons en déduire que dans un contexte d'enregistrement bruité et hétérogène, dans lequel l'apprentissage automatique est difficile, cette approche « bas niveau » est pertinente. Cette approche générique produit des résultats exploitables, quel que soit le type d'enregistrement et sans connaissances a priori. Dans des conditions ciblées, il peut être intéressant d'effectuer une méthode fondée sur l'apprentissage si nous avons à notre disposition des données annotées. Au final, il semble préférable d'utiliser les méthodes dites classiques dans un second temps, après avoir effectué une première segmentation qui réduit considérablement l'hétérogénéité du corpus. Cette approche fait écho à d'autres travaux effectués sur des données hétérogènes, où une première segmentation basée sur des descripteurs choisis spécifiquement précède une classification par des méthodes classiques afin d'augmenter le taux de reconnaissance. Elle a par exemple été utilisée pour la détection d'impacts de balles dans des vidéos sportives [ZDC06].

Approche granulaire

Une autre approche du phénomène des sons d'eau a été développée dans le chapitre 4. À partir d'un modèle physique, nous avons alors proposé la détection d'événements discrets localisés dans le plan temps/fréquence, correspondants aux vibrations de bulles d'air dans l'eau. Si nous revenons sur la genèse de cette approche, nous pouvons observer que l'analyse seule des spectrogrammes n'a pas suffi à comprendre le phénomène pour pouvoir le modéliser entièrement. Ce sont les recherches en synthèse sonore qui nous ont permis d'identifier des propriétés caractéristiques de ces sons.

Cette approche semble d'autant plus intéressante que certaines études de synthèse se focalisent sur l'aspect réaliste du son produit plutôt que sur la modélisation complète du système physique. Ces études s'attachent à trouver des propriétés d'un son produit qui le rendent perceptivement réaliste [MAYKM13], et éventuellement saillant dans un contexte bruité. Ainsi pour les sons d'eau, les interactions solides ne sont pas modélisées dans les études, ce qui simplifie les modèles physiques utilisés. L'utilisation de ces modèles est ainsi plus accessible et efficace pour

la génération de son, mais peut être aussi pour leur détection. Ainsi, la reconnaissance automatique de sons environnementaux à partir des algorithmes de synthèse sonore pourrait être une approche intéressante pour d'autres types de recherches.

Par ailleurs, nous pouvons également revenir sur notre méthode détection des bulles d'air en vibration. Pour cette tâche, il semble qu'une approche basée des fenêtres temporelles courtes soit mal adaptée. En effet la détection de zones temps/fréquence précises nécessite des fenêtres temporelles assez larges, de l'ordre de 300 ms par exemple pour les gouttes d'eau. D'une part, les systèmes classiques, par exemple des GMM ou des SVM utilisant une paramétrisation sur des courtes fenêtres, semblent inadaptés à cette problématique. D'autre part, un système basé sur des HMM ne semble pas non plus approprié, car les évènements sonores produits par les bulles d'air arrivent simultanément, et non successivement.

Une approche par décomposition en atomes, telle la décomposition en matrices non négatives [Ber09] ou en atomes de Gabor [CNK09], pourrait être plus adaptée à cette détection d'évènements simultanés. Des adaptations pourraient être envisagées pour prendre en compte les similarités entre les atomes à différentes fréquences, telle celle effectuée dans [GGWM11], où les mêmes atomes sont étirés temporellement en fonction de la fréquence. Toutefois, la diversité des types de vibrations, ainsi que l'aspect bruité de notre corpus de travail rend difficile l'application directe de ces méthodes. Ce type d'approche reste néanmoins une perspective intéressante pour les enregistrements effectués dans des conditions maîtrisées.

D'un point de vue théorique, nous pourrions comparer cette décomposition du signal en atomes à la technique de synthèse correspondante, la *synthèse granulaire* [Roa88]. La factorisation de ces atomes, ou grains, correspond donc à une tâche d'*analyse granulaire*. Ce cadre théorique semble intéressant et pourrait être appliqué à la détection d'autres types de sons bruités, comme les sons de frottement ou de grattement. Ces recherches pourraient également être appliquées à d'autres types de sons pouvant être synthétisés par synthèse granulaire.

Perception des sons environnementaux

Le chapitre 5 est dédié à la reconnaissance perceptive des sons de liquide. Ce travail constitue à notre connaissance la première expérience de catégorisation effectuée exclusivement sur un corpus de sons de liquide. Ainsi, si d'autres expériences pourraient éventuellement permettre de valider ou de préciser les catégories obtenues, ce premier résultat constitue une base de travail notable.

Dans la lignée de la taxonomie des sons environnementaux de Gaver, souvent utilisée comme référence, il est intéressant de considérer les résultats obtenus dans le contexte plus large de l'écoute des sons environnementaux. Nos conclusions s'additionnent aux résultats d'études de perception obtenus sur d'autres types de corpus; notamment les expériences de [HLM$^+$12] effectuées sur un ensemble varié de sons environnementaux, et sur un ensemble de sons de solides. Ces différentes études sont complémentaires et peuvent permettre d'obtenir une vision globale d'une représentation type du monde sonore. La figure 3 permet de visualiser un dendrogramme global des sons environnementaux à partir des résultats de ces trois expériences.

Nous avons décrit au chapitre 5 des différences entre nos résultats et la taxonomie de Gaver. En particulier, les classes de sons hybrides proposées par Gaver (liquide/solide, liquide/gaz, liquide/solide/gaz) sont ici groupées parmi les sons de liquide. Nous pouvons considérer que la classe hybride liquide/solide correspond aux classes : *filet sur surface, jet, remplir, évacuation*. De plus, comme nous l'apprennent les modèles acoustiques, la présence de l'air est souvent nécessaire pour que le son d'eau soit audible. Nous pouvons donc considérer que toutes les classes de sons de liquides (à part peut être les classes *jet* et *filet sur surface*), sont des classes hybrides faisant

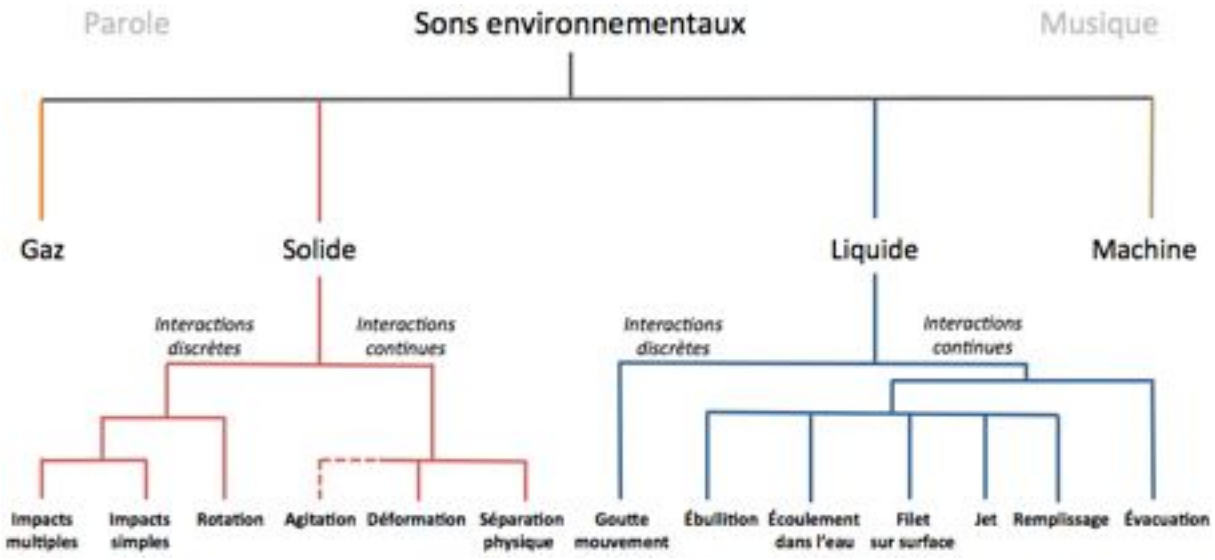


FIGURE 3 – Taxonomie des sons environnementaux.

intervenir l'élément gazeux.

Par ailleurs, les interactions de type discret et continu se retrouvent dans les classes de solides et de liquides. À la différence des sons de solides, les catégories obtenues à partir des sons de liquide ne sont pas équilibrées dans ces deux types d'interactions. Les corpus utilisés dans les expériences pourrait être à l'origine de cette différence.

Nous pouvons également remarquer qu'à la différence des sons d'objets solides, les sons de liquide sont plutôt étiquetés par des labels n'impliquant pas de geste. De manière générale, nous pouvons supposer que les sons de solides sont plus souvent provoqués par des gestes que les sons de liquides.

La taxonomie obtenue permet au final de préciser la manière dont les personnes se représentent le monde sonore. Ces expériences de perception peuvent avoir des applications dans d'autres domaines, par exemple le *design sonore*, dans lequel l'étude de la manière dont les personnes écoutent et organisent les sons est fondamentale.

Pour compléter cette taxonomie, d'autres classes pourraient être étudiés par la même approche, comme les classes de sons produits par des interactions aérodynamiques, ou celle des évènements sonores produits par des machines.

Annexe A

Évaluation des compétences praxiques lors d'activités de la vie quotidienne

Entretien corporel / Hygiène		
Activités	Compétences évaluées	
Se laver les dents	Reconnaissance	Brosse à dents Dentifrice
	Utilisation / coordination	Porter à la bouche Rinçage
S'habiller	Reconnaissance	Choisit ses vêtements Propreté / saison
	Utilisation / coordination	Orientation au corps Enfilage Laçage / zip
Alimentation		
Activités	Compétences évaluées	
Tartiner	Reconnaissance	Couteau / cuillère
	Utilisation / coordination	Manche / lame Ordre des séquences
Faire la cuisine	Reconnaissance	Ingrédients Outils
	Utilisation / coordination	Choix des recettes Préparation
Utiliser une machine (café, bouilloire, toaster)	Reconnaissance	Choix des accessoires (thé, café, filtres)
	Utilisation / coordination	Programmation des séquences Chauffe / réceptacle
Utiliser une machine (lave-lige / micro-onde)	Reconnaissance	Accessoires
	Utilisation / coordination	Programme

Entretien du domicile		
Activités	Compétences évaluées	
Laver la vaisselle (main)	Reconnaissance	Produits / éponge
	Utilisation / coordination	Lavage
Essuyer la vaisselle	Reconnaissance	Torchon
	Utilisation / coordination	Torchon
Passer le balais / Aspirateur	Reconnaissance	Matériel
	Utilisation / coordination	Tenu du balais Adapté à la zone
Distractions (seul)		
Activités	Compétences évaluées	
Regarder la télévision / Écouter la radio	Reconnaissance	Appareil Matériel
	Utilisation / coordination	Allumer le poste Choix du programme Navigation
Tricot / broderie	Reconnaissance	Aiguilles / fil
	Utilisation / coordination	Aiguilles / fil Respect du modèle
Bricolage	Reconnaissance	Outils
	Utilisation / coordination	Outils Résulat
Jardinage	Reconnaissance	Outils
	Utilisation / coordination	Outils Résulat
Relations sociales		
Activités	Compétences évaluées	
Servir un plat / verre	Reconnaissance	Matériel
	Utilisation / coordination	Réceptacle Matériel
Téléphoner	Reconnaissance	Recherche du numéro
	Utilisation / coordination	Composition du numéro
Répondre au téléphone	Reconnaissance	Identifie la sonnerie
	Utilisation / coordination	Décrochage Conversation
Ouvrir la porte	Reconnaissance	Identifie la sonnerie
	Utilisation / coordination	Orientation / déplacement Accueil

Annexe B

Le système auditif humain

Le système auditif permet d'analyser les sons de manière temporelle et fréquentielle. Le schéma général de l'appareil auditif est illustré à la figure B.1.

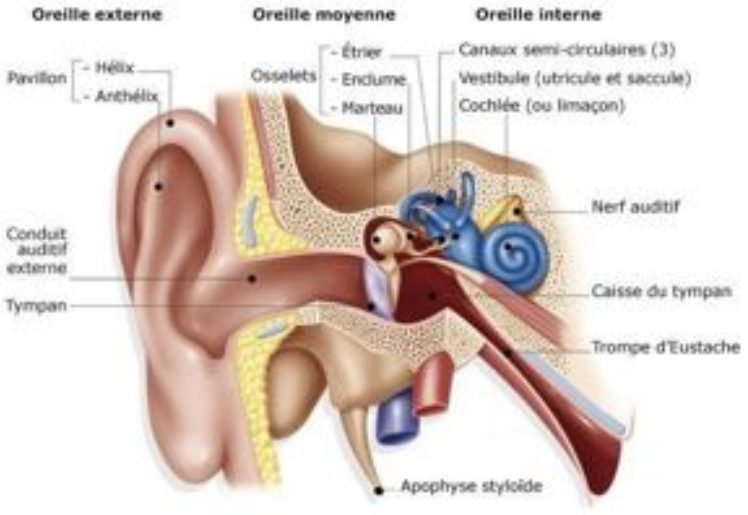


FIGURE B.1 – Appareil auditif humain.

L'oreille externe

L'oreille externe permet un filtrage fréquentiel. Un des résultat est l'amplification des fréquences autour de 3kHz.

L'oreille moyenne

L'oreille moyenne est composé du tympan et de la chaîne d'osselets. Le tympan transmet les vibrations acoustiques aux osselets. Deux muscles, celui du marteau et de l'étrier, permettent de modifier l'amplitude de vibration. Le muscle de l'étrier assure ainsi une fonction de protection importante.

L'oreille interne

L'organe vestibulaire et la cochlée constituent l'oreille interne. La cochlée est un canal osseux en forme de spirale qui permet la transduction de la vibration. Elle est responsable de la perception fréquentielle, selon un processus passif (onde propagée de Békésy) ou actif [ADD$^+$88].

La modélisation de l'appareil auditif humain a lancé de nombreux travaux de recherche. Certains procédés de représentation du signal sous la forme temps/fréquence sont ainsi clairement inspirés des traitements effectués par le système auditif humain, comme par exemple les échelles de Mel, les gammatones (voir annexe C), et différents types de cochléogrammes [BDVA$^+$08, LKD10].

Annexe C

Descripteurs acoustiques

C.1 Les paramètres temporels

C.1.1 Énergie

L'énergie est probablement le descripteur le plus couramment utilisé. C'est aussi un descripteur sonore bien appréhendé par les non-spécialistes, car la notion d'énergie ou de puissance du signal est couramment évoquée, par exemple lors de concert, en écologie sonore, ou dans le cas de la protection auditive.

Soit x_n un signal numérique discret à support fini avec $n \in [1, N]$.

L'énergie, ou valeur quadratique du signal, est définie par :

$$E = \sum_{n=1}^{N} x_n^2 \tag{C.1}$$

L'énergie à court terme est généralement calculée sur de petites fenêtres du signal.

C.1.2 Le ZCR

Le ZCR ou Zero Crossing Rate représente le nombre de passages à zéro du signal dans une trame. Les sons périodiques ont tendance à avoir une faible valeur du ZCR, alors que les sons bruités ont plutôt tendance à avoir des valeurs élevées. Ce paramètre a été utilisé dans la détection d'activité vocale [SS97], ou la détection de sons percussifs [GPD00].

Voici l'équation du ZCR de la trame i du signal :

$$ZCR = \frac{1}{2N} \sum_{n=1}^{N} (|sign(x_n) - sign(x_{n-1})|) \tag{C.2}$$

Avec x_n le n$^{\text{ième}}$ échantillon de la trame i et N le nombre d'échantillon dans la trame i.

C.2 Les paramètres spectraux

La transformée de Fourier discrète d'un signal discret s'écrit :

$$w(k) = \sum_{n=0}^{N-1} x(n).e^{-2i\pi k \frac{n}{N}} \qquad pour k \in [0, N] \tag{C.3}$$

La partie réelle de cette transformée de Fourier nous donne les coefficients du spectre, qui constituent la matière première de l'analyse spectrale. Ils peuvent ainsi être utilisés directement comme descripteurs du son (bien que les MFCC soient plus populaires). Une analyse par fenêtre successive permet d'observer l'évolution des coefficients spectraux dans un spectrogramme. Par contre, la transformée de Fourier d'un signal fenêtré produit des artefacts que des fenêtrages particuliers permettent de minimiser, par exemple les fenêtres de Hamming.

Les paramètres suivants, dit paramètres spectraux bas niveaux, sont directement calculés à partir des coefficients de la transformée de Fourier. Ils peuvent en outre être calculés par bandes de fréquence, ce qui permet de les utiliser de manière plus précise.

C.2.1 Centroïde spectral

Le centroïde spectral, ou centre de gravité spectral (CSG), est un descripteur très utilisé. Il révèle l'aspect grave ou aigu d'un son. Il correspond au barycentre des fréquences dont le poids serait l'amplitude. Voici l'équation du centre de gravité spectral de la trame i:

$$CSG = \frac{\sum_{n=1}^{N} f(k)w(k)}{\sum_{n=1}^{N} w(k)} \tag{C.4}$$

où $w(k)$ est l'amplitude de la fréquence $f(k)$.

Plusieurs autres paramètres dérivent du centroïde, dont le *spectral spread*, ou étalement spectral, qui correspond à la variance du spectre (au moment d'ordre 2). Cette mesure rend compte de l'étalement de l'énergie fréquentielle autour du centroïde.

$$Spread = \frac{\sum_{n=1}^{N} (f(k) - CSG)^2 w(k)}{\sum_{n=1}^{N} w(k)} \tag{C.5}$$

Le *Spectral Skewness* correspond au moment d'ordre 3 du spectre d'énergie, et il donne un indice quant à la symétrie de la distribution autour du centroïde.

$$Skewness = \frac{\sum_{n=1}^{N} (f(k) - CSG)^3 w(k)}{\sum_{n=1}^{N} w(k)} \tag{C.6}$$

Le *Spectral Kurtosis* est le moment d'ordre 4 du signal. Ce paramètre donne une mesure de la platitude du spectre autour du centroïde.

$$Kurtosis = \frac{\sum_{n=1}^{N} (f(k) - CSG)^4 w(k)}{\sum_{n=1}^{N} w(k)} \tag{C.7}$$

C.2.2 Le *spectral rollof*

Le spectral rollof est une autre mesure utilisée pour quantifier la position de l'énergie spectrale. Cette mesure correspond pour une trame à la fréquence en dessous de laquelle se trouve une proportion e, en général 95%, de l'énergie.

Pour calculer la fréquence de spectral rollof $f(l)$, il faut donc trouver l tel que

$$\sum_{n=1}^{l} w(k) > e \quad et \quad \sum_{n=1}^{l-1} w(k) \leq e \quad avec \quad e = 0.95 \sum_{n=1}^{N} w(k) \tag{C.8}$$

Le flux spectral

Le flux spectral mesure la variation des coefficients spectraux entre deux trames consécutives. La formule suivante permet de calculer le flux spectral entre la trame i et la trame $i-1$ du signal.

$$FS = \sum_{n=1}^{N} \left(\frac{w_i(k)}{\sum_{n=1}^{N} w_i(k)} - \frac{w_{i-1}(k)}{\sum_{n=1}^{N} w_{i-1}(k)} \right)^2 \tag{C.9}$$

C.2.3 Le *spectral slope*

Le spectral slope calcule la pente du spectre. Il est calculé par régression linéaire sur les coefficients spectraux.

$$slope = \frac{N \times \sum_{n=1}^{N} f(k)w(k) - \sum_{n=1}^{N} f(k) \times \sum_{n=1}^{N} w(k)}{N \times \sum_{n=1}^{N} f^2(k) - (\sum_{n=1}^{N} f(k))^2} \tag{C.10}$$

La figure C.1 illustre la pente du spectre calculé sur une trame du signal.

C.2.4 Le *spectral flatness*

Le spectral flatness est calculé comme la moyenne géométrique du spectre d'énergie, divisée par sa moyenne arithmétique.

$$SF = \frac{\sqrt[N]{\prod_{k=1}^{N} w(k)}}{\frac{1}{N} \sum_{i=1}^{N} w(k)} \tag{C.11}$$

C.2.5 La fréquence fondamentale

Si un signal sonore est périodique ou pseudo-périodique sur une fenêtre donnée, avec T une période, nous avons :

$$x(t) = x(t+T), \forall t \tag{C.12}$$

Alors ce signal évoquera un pitch qui varie avec sa fréquence fondamentale $F_0 = \frac{1}{T}$.

Dans les sons considérés comme harmoniques, tels les voyelles dans la voix parlée, les sons des instruments à vent, ou les instruments à corde frottées, les fréquences présentes dans le spectre

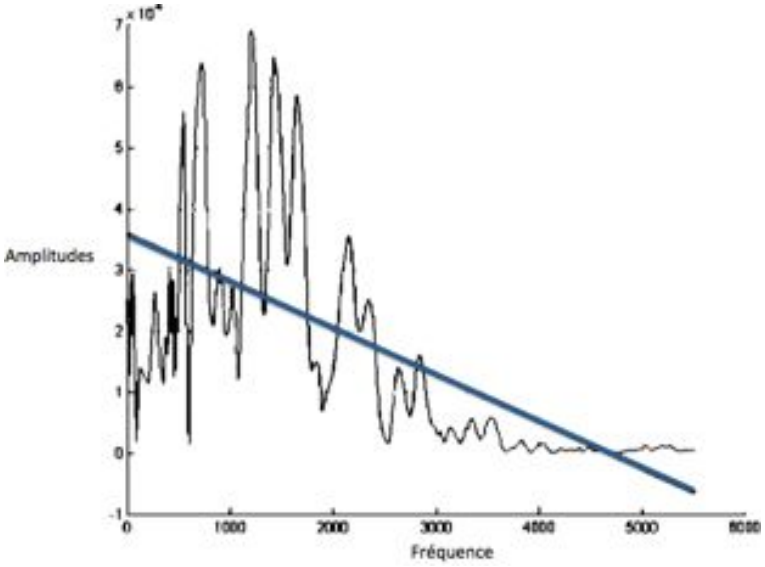


FIGURE C.1 – Pente du spectre.

sont toutes des multiples entiers de cette fréquence fondamentale F_0. La figure C.2 nous montre une représentation d'un son harmonique avec sa fréquence fondamentale F_0 et ses partiels.

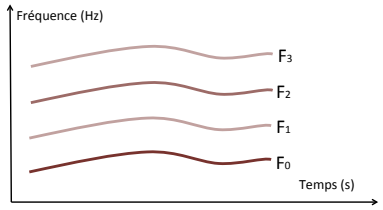


FIGURE C.2 – Représentation temps/fréquence d'un son harmonique.

Il existe une littérature très dense sur la fréquence fondamentale, notamment sur sa perception, ainsi que le pitch dans les sons inharmoniques. La tache d'extraction automatique de fréquence fondamentale a motivé de nombreux efforts au cours de ces dernières décennies [Hes83]. L'extraction de fréquence fondamentale dans le cas de signaux à plusieurs sources continue d'être un domaine de recherche très actif [KD06].

De nombreuses méthodes ont été proposées, dans le domaine temporel ou spectral. Une méthode classique pour estimer la fréquence fondamentale dans le domaine temporel consiste à calculer la fonction d'autocorrélation:

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x(j)x(j+\tau) \tag{C.13}$$

où τ est le décalage temporel et W la fenêtre d'intégration.
Ce calcul permet de vérifier la similarité du signal par rapport à des fenêtres τ de taille variable. Le calcul de F_0 à partir de cette fonction d'autocorrélation se fait en général en choisissant le plus

grand pic différent de zéro. Cette méthode est assez simple, mais génère de nombreuses erreurs, en particulier des erreurs d'estimation dues à des fenêtre τ trop petites ou trop grandes.

Dans cette thèse, nous avons utilisé un algorithme célèbre, dérivé de cette fonction d'autocorrélation [DCK02]. Le Yin permet de détecter la fréquence fondamentale de manière assez robuste et efficace. Il utilise une variante de la fonction d'autocorrélation, soit :

$$d_t(\tau) = \sum_{j=t+1}^{t+W} [x(j) - x(j+\tau)]^2 \tag{C.14}$$

et permet d'éviter de nombreuses erreurs d'estimation, dues notamment à des modulations d'énergie importantes dans le signal.

C.2.6 Les coefficients cepstraux

Les coefficients cepstraux ou MFCC pour *Mel frequency cepstral coefficients* constituent probablement le descripteur le plus utilisé en reconnaissance automatique audio. Le calcul de coefficients cepstraux suit un procédé en plusieurs étapes :

- calcul des coefficients de Fourier,
- filtrage à travers l'échelle Mel,
- logarithme des coefficients filtrés,
- transformée en cosinus discrète.

La formulation peut légèrement varier, en particulier la transformée en cosinus discrète est parfois remplacée par une transformée de Fourier inverse. Étant donné leur utilisation massive et leur aptitude à modéliser les signaux de parole, il existe une littérature importante sur les MFCC. Il semblerait que malgré leur utilisation importante dans le domaine des sons environnementaux, ils ne soient pas très robustes au bruit [HLDC+04].

C.2.7 Les gammatones

Les *Gammatones Cepstral Coefficients*, ou GTCC, que nous appellerons gammatones, constituent un descripteur biologiquement inspiré qui est très souvent utilisé dans le domaine CASA.

Les gammatones sont assez proches des MFCC au niveau computationnel, car l'unique différence avec le procédé de calcul des MFCC décrit ci-dessus réside dans le filtrage, qui est dans le cas des gammatones effectué par les filtres gammatones dont la réponse impulsionnelle est le produit d'une distribution gamma avec une sinusoïde centrée sur la fréquence f_c:

$$g(t) = Kt^{(n-1)}e^{-2\pi Bt}cos(2\pi f_c t + \phi) \qquad t > 0 \tag{C.15}$$

où K est l'amplitude, n l'ordre du filtre, ϕ la phase et B la durée de la réponse impulsionnelle.

Un paramètre entrant dans le calcul des filtres gammatones est la largeur de bande rectangulaire équivalente ou *Equivalent Rectangular Bandwith* (ERB), qui donne une approximation de la bande passante à chaque point de la cochlée. Les ERB sont utilisés dans l'implémentation des gammatones pour calculer la distance entre filtres consécutifs.

Concrètement, les filtres gammatones ainsi constitués sont en général plus larges que les filtres issus de l'échelle Mel utilisés pour les MFCC. Le recouvrement entre chaque filtre est également plus important. Au niveau des résultats, les gammatones présentent une alternative intéressante aux MFCC. Il semblerait qu'ils soient mieux adaptés à la détection de sons environnementaux que les MFCC [VA12].

La représentation temps/fréquence associé aux filtres gammatones est en général appelée cochléogramme.

C.2.8 Le coefficient de variation

D'autres paramètres peuvent être produits de façon *ad hoc*. Par exemple, Wilson propose d'utiliser le coefficient de variation du spectre comme mesure du bruit.

Le coefficient de variation est une mesure de dispersion relative. Il a été utilisé pour caractériser l'aspect bruité d'un signal [WIL94]. Le coefficient de variation est calculé comme un rapport entre l'écart type σ et la moyenne du spectre d'énergie du signal μ.

$$Y = \frac{\sigma}{\mu} \tag{C.16}$$

Ce coefficient est utilisé pour déterminer si le signal est proche d'un bruit blanc. En effet, les coefficients de la transformée de Fourier d'un bruit blanc sont des variables indépendantes dont la moyenne est égale à leur écart type. Dans le cas d'un bruit blanc, le coefficient de variation vaudra donc 1.

C.2.9 Autres paramètres

Nous avons présenté quelques paramètres utilisés dans ce document. Il faut souligner l'existence de nombreux autres paramètres, dont nous ne détaillerons pas le calcul. Par exemple, les coefficients de prédiction linéaire (LPC) ou de prédiction linéaire perceptive (PLP), souvent utilisés dans le cas de la parole [LSDM01, Her90].

Annexe D

Modèles de classification

Dans cette partie nous allons présenter quelques méthodes statistiques couramment utilisées pour résoudre un problème de détection ou de classification automatique.

D.1 Les k-plus proches voisins

L'idée de cet algorithme est de considérer les k voisins les plus proches de l'observation x. Nous pouvons représenter cette idée de k plus proches voisins sur un plan. Par exemple, sur la figure D.1, on considère dans un espace à 2 dimensions les 6 voisins les plus proches de l'observation. Selon l'algorithme, cette observation sera attribuée à la classe majoritaire, ici les triangles verts.

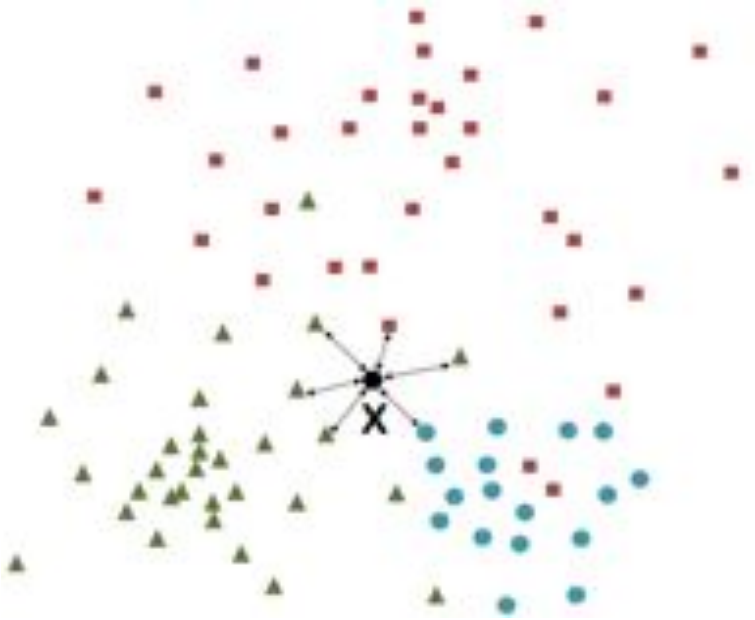


FIGURE D.1 – Les 6 plus proches voisins autour de x.

La distance utilisée pour trouver les voisins les plus proches a évidemment un impact déterminant sur les résultats de l'algorithme. La distance la plus couramment utilisée est la distance euclidienne.

Une des difficultés en apprentissage automatique est de choisir le nombre k. Le choix de la distance est également non trivial, des poids peuvent par exemple être appliqués à chaque dimension lors du calcul de la distance, ce qui permet en pratique de renforcer l'influence d'un descripteur. En pratique un ensemble de validation est utilisé pour affiner ces choix.

D.2 Les modèles de mélange de lois gaussiennes

Pour un espace de données de dimensions D, chaque loi gaussienne m d'un mélange à M composantes est paramétrée par un vecteur de moyenne μ_m de dimension D et une matrice de covariance Σ_m de dimension $D \times D$. Chacune de ces lois est également pondérée par une probabilité *a priori* w_m, avec $\Sigma_m^M w_m = 1$.

Ainsi, si y est un vecteur d'observation de $\mathbf{R}^D$, la probabilité d'observer y dans la classe C_i s'écrit:

$$P(y|C_i) = \sum_{m=1}^{M} w_m \mathcal{N}(y|\mu_m, \Sigma_m), \tag{D.1}$$

où $\mathcal{N}$ est une loi normale multidimensionnelle:

$$\mathcal{N}(y|\mu_m, \Sigma_m) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} e^{-\frac{1}{2}(y-\mu_m)^t \Sigma_m^{-1}(y-\mu_m)} \tag{D.2}$$

L'utilisation de cette méthode en reconnaissance automatique nécessite dans un premier temps l'apprentissage des paramètres de ces lois gaussiennes. En général, ces paramètres sont optimisés par la méthode du *maximum de vraisemblance*. Cette procédure se fait le plus souvent itérativement via l'algorithme *espérance-maximisation* (EM) [DLR77].

D.3 Les machines à vecteur de support

D.3.1 Méthode

Les machines à vecteurs de support ou SVM pour *Support Vector Machine* sont devenues en quelques années des méthode de classification incontournables. Sur des problèmes de classification audio, elles semblent en général donner des résultats légèrement supérieurs aux GMM. Elles permettent de classifier des échantillons en deux classes.

Pour déterminer la frontière, ou l'hyperplan optimal, qui sépare deux classes, les SVM s'appuient sur les échantillons les plus proches de la frontière, que l'on appelle *vecteurs supports*. Pour s'élargir aux problèmes non linéairement séparables, la méthode des SVM propose de transformer l'espace des données en un espace de plus grande dimension, afin de trouver une séparation linéaire.

Soit un ensemble d'apprentissage (x_n, y_n) ou y_i est le label de l'échantillon x_i qui peut prendre deux valeurs, $y_i = \pm 1$.

Dans le cas linéaire, le problème consiste à trouver l'hyperplan séparant au mieux les données (voir figure D.2).

Un classifieur linéaire s'obtient par un vecteur de poids $w = (w_1, ..., w_n)$ et un biais w_0.

$$h(x) = w^T x + w_0 \tag{D.3}$$

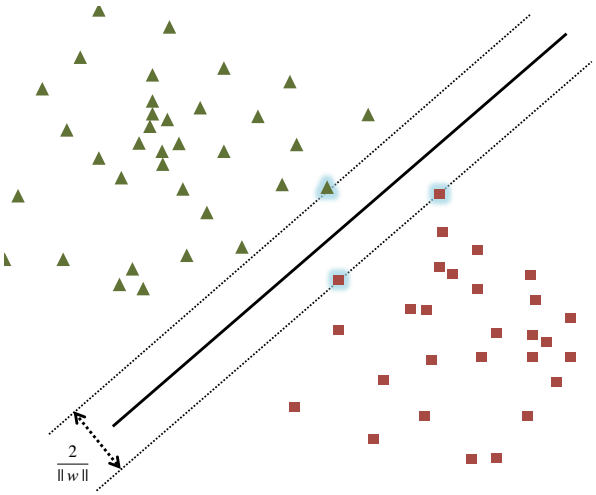


FIGURE D.2 – Hyperplan séparateur dans un cas linéairement séparable.

Nous introduisons deux hyperplans parallèles au premier de façon à ce qu'il n'y ait pas d'échantillons dans la marge comprise entre ces deux hyperplans. Le problème revient à maximiser la marge entre deux hyperplans parallèles, donnée par $\frac{2}{||w||}$.

On utilise en général le multiplicateur Lagrangien pour résoudre ce problème d'optimisation.

Dans le cas non linéaire (voir figure D.3), l'idée est de transposer le problème dans un espace de plus grande dimension, éventuellement infini, dans lequel il deviendra linéaire. On applique ainsi à x une transformation non-linéaire ϕ.

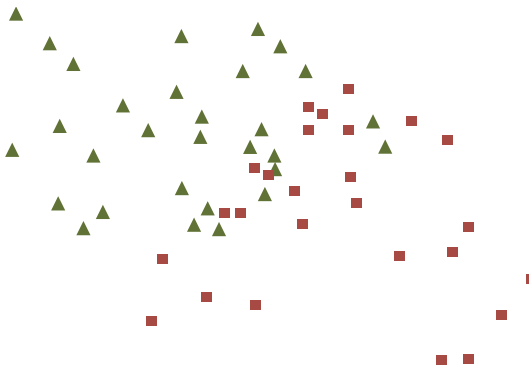


FIGURE D.3 – Cas non linéairement séparable.

On cherche alors l'hyperplan:

$$h(x) = w^T \phi(x) + w_0 \tag{D.4}$$

Pour résoudre le problème d'optimisation, on utilise une fonction noyau qui vérifie

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{D.5}$$

En pratique, il existe quatre fonctions noyaux de base: noyau linéaire, noyau polynomial, noyau Radial Basis Function (RBF), et noyau sigmoïde [Bur98].

D.3.2 Extension du domaine à plus de deux classes

En pratique, les SVM sont couramment utilisés pour des problèmes de classification multi-classes. Pour une classification à N classes, une approche courante consiste à utiliser N classifieurs SVM. Le i$^{\text{ème}}$ SVM est dans ce cas entraîné avec les échantillons de la i$^{\text{ème}}$ classe avec des labels positifs, et tous les autres échantillons avec des labels négatifs. Le résultat de ce système à N-SVM est la classe correspondant au SVM ayant la plus grande valeur [Vap98].

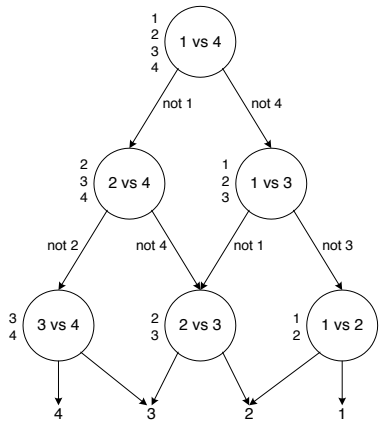


FIGURE D.4 – Graphe orienté acyclique pour résoudre un problème de classification multi-classe.

Une autre approche consiste à étudier chaque combinaison de deux classes, et de construire C_N^k classifieurs SVM. Dans cette approche, il n'est toutefois pas toujours facile d'avoir une méthode fiable pour déterminer le résultat du classifieur global. Une des solutions peut être la construction d'un graphe orienté acyclique (Directed Acyclic Graph ou DAG), comme illustré à la figure D.4 [PCST99].

D.4 Les modèles de Markov Cachés

Une chaîne de Markov est un modèle statistique composé d'états et de transitions. Chaque transition est associée à une probabilité d'être emprunté. La figure D.5 illustre une chaîne de Markov.

Dans un modèle de Markov caché, l'état n'est pas directement observable. Une séquence d'observation $e_1, ..., e_M$ est issue d'un alphabet de taille M et dépend d'une séquence d'observations. Une matrice de probabilités contient la probabilité $b_i(e_k)$ pour chaque symbole e_k d'être émis par un état q_i.

$$b_i(e_k) = P(e_k(t)|s_t = q_i) \tag{D.6}$$

avec

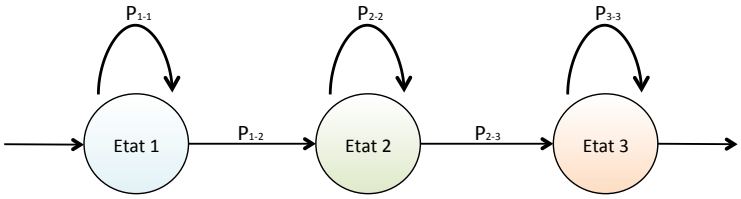


FIGURE D.5 – Chaine de Markov à 3 états.

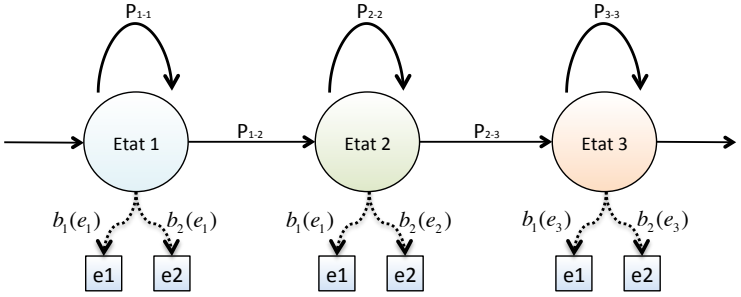


FIGURE D.6 – Modèle de markov caché à 3 états, avec un alphabet de taille 2.

$$\sum_{k=1}^{M} b_i(e_k) = 1 \tag{D.7}$$

La figure D.6 illustre un modèle de Markov à trois états avec un alphabet d'observations de taille 2.

Pour utiliser un modèle MMC, deux problèmes doivent être résolus :

- estimer les paramètres du modèle qui maximise la probabilité d'observation de séquences. Ce problème est en général résolu en utilisant l'algorithme Expectation-Maximisation.
- trouver le meilleur trajet dans une séquence d'état à partir d'une séquence d'observation. Ce problème est en général résolu grâce à l'algorithme de Viterbi [Vit67].

D.4.1 La factorisation en matrices non négatives

Soit V une matrice de dimensions $F \times N$ à coefficients réels positifs ou nuls. Le but de la NMF et de l'approximer par deux matrices W et H tel que :

$$V \approx WH \tag{D.8}$$

Dans le cas optimum, la matrice W représente ainsi les spectres de chacune des sources, et la matrice H leur activation. La figure D.7 nous montre un exemple de décomposition d'un spectrogramme selon les sources et leur activation. Dans une décomposition NMF, les colonnes de la matrice W pourraient être les trois spectres de chacune des notes (à gauche). De même les lignes de matrice H pourraient correspondre aux activations de chacune de ces trois notes (en haut). Le produit de ces deux matrices donnant le spectrogramme (en bas à droite).

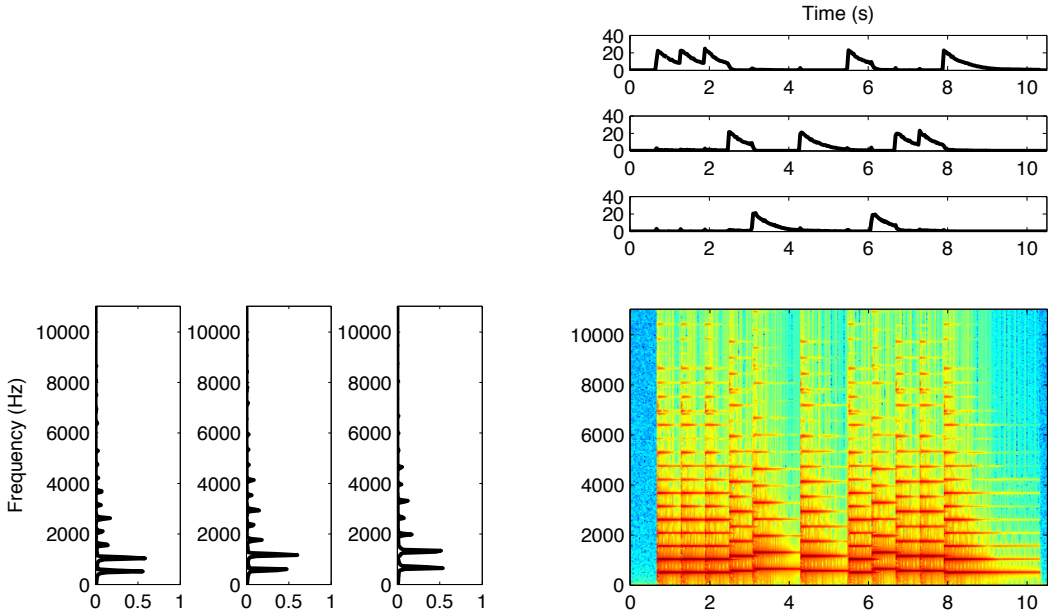


FIGURE D.7 – NMF de l'air « Au clair de la lune » joué sur un piano synthétique [Ber09].

Appliquer une NMF revient à déterminer W et H. Il existe pour cela plusieurs méthodes itératives, supervisée ou non supervisée. Le problème peut être résolu en minimisant une fonction de coût F :

$$F(W, H) = ||V - WH||^2 \tag{D.9}$$

Plusieurs méthodes ont été développées pour résoudre ce problème. L'algorithme de Lee et Seung est assez utilisé pour sa simplicité d'implémentation [SL01]. Cependant, la plupart de ces méthodes se heurtent à la détection d'un minimum local dans le processus itératif.

D.4.2 Autres méthodes

D'autres modèles statistiques sont régulièrement utilisés en classification. Si nous n'allons pas détailler ici leur principes, nous pouvons citer les réseaux de neurones [Bis95], ou encore les algorithmes de *boosting* [GZL01].

Annexe E

Sons utilisés lors de l'expérience perceptive

E.1 Corpus

Dans cette section, nous présentons les sons utilisés lors des expériences perceptives du chapitre 5. Chaque son est identifié par son numéro *n*. Nous présentons pour chaque stimulus son label et sa durée en seconde.

Une croix dans la colonne *Ident* signifie que le son a été mal identifié dans l'expérience 2 et qu'il a été par la suite supprimé du corpus. Ces sons n'apparaissent donc pas dans l'expérience 3 de catégorisation libre.

La dernière colonne présente l'indice d'origine des sons permettant de l'associer aux fichiers des bases de données sonores utilisées.

n	Label	Durée	Ident	*Ind*
1	**Remplir un lavabo**	4.99		# 1
2	**Vider un lavabo**	2.67		# 2
3	**Gouttes tombant dans un évier**	1.78		# 3
4	**Agiter un pommeau de douche**	4.14		# 4
5	**Jet de douche**	4.1		# 5
6	**Robinet grand ouvert**	2.25		# 6
7	**Gouttes dans un gobelet en plastique**	1.32		# 7
8	**Remplir un gobelet en plastique**	3.71		# 8
9	**Verser de la peinture dans un pôt**	4.54		# 9
10	**Verser dans un bac en plastique**	2.9		#10
11	**Verser dans un récipient vide en plastique**	3.56		#11
12	**Verser dans un vase vide**	3.72		#12
13	**Secouer une bouteille en plastique remplie**	2.02		#13
14	**Secouer un vase contenant de l'eau**	2.23		#14
	Secouer un bidon en plastique rempli	1.01	**X**	#15
	Secouer un bidon contenant du liquide et des solides	3.05	**X**	#16
	Secouer un gobelet en plastique	2.03	**X**	#17
	Eclaboussures sur du carton	0.39	**X**	#18
	Mélanger de la peinture	1.45	**X**	#19
	Eclaboussures sur une plaque de métal	1.65		#20
15	**Verser sur une plaque de métal**	2.93		#21

n	Label	Durée	Ident	*Ind*
16	**Agiter de l'eau avec ses doigts**	0.63		#22
	Eclaboussures sur du plastique	0.82	**X**	#23
17	**Presser une éponge au dessus de l'évier**	3.96		#24
	Presser une éponge	1.6	**X**	#25
18	**Remplir une théière**	5.04		#26
19	**Faire frire un œuf**	2.51		#27
20	**Tirer la chasse d'eau**	4.49		#28
21	**Bouger dans un bain**	2.64		#29
22	**Objet tombant dans le bain**	1.19		#30
23	**Bulles dans l'eau**	0.99		#31
24	**Ebullition dans une grande casserole**	4.08		#32
25	**Souffler avec une paille à la surface de l'eau**	2.55		#33
26	**Secouer une bouteille en plastique**	3.55		#34
	Presser une matière visqueuse	1.38	**X**	#35
	Impact de matière visqueuse	1.26	**X**	#36
27	**Vider une bouteille**	3.85		#37
28	**Gouttes dans un évier**	5.43		#38
29	**Flot d'eau dans un grand évier**	3.21		#39
30	**Remplir un bidon en plastique**	5.25		#40
31	**Remplir un évier**	3.09		#41
	Gouttes tombant sur des feuilles	2.53	**X**	#42
32	**Goutte tombant dans l'eau**	1.2		#43
33	**Agiter un récipient en métal à moitié rempli**	4.86		#44
34	**Essorer des habits mouillés**	1.69		#45
35	**Pluie tombant sur le sol depuis le toit**	1.96		#46
36	**Ebullition dans un chaudron**	4.56		#47
37	**Mettre un glaçon dans une boisson**	2.55		#48
38	**Pétillement après ouverture d'une cannette**	2.87		#49
39	**Remplir un verre contenant des glaçons**	2.21		#50
40	**Agiter un récipient en métal à moitié rempli**	2.81		#51
41	**Vider un lavabo**	5.01		#52
42	**Jet de douche dans une baignoire**	5.12		#53
	Ouvrir un robinet	5.01	**X**	#54
	Huile de friture bouillante	1.85	**X**	#55
43	**Steack frit sur un grill**	3		#56
44	**Eau bouillante dans une casserole**	4.26		#57
45	**Vider un évier**	4.73		#58
46	**Jet de douche avec le bruit d'évacution**	4.13		#59
47	**Remplir un évier**	3.65		#60
	Verser du ketchup	3.53	**X**	#61
	Pulvériser de l'eau	1.38	**X**	#62
48	**Pulvériser dans un verre**	1.48		#63
49	**Verser de la bière dans un verre**	4.69		#64
50	**Agiter un verre contenant des glaçons**	2.66		#65
51	**Remplir une baignoire**	3.4		#66
52	**Eau coulant dans baignoire partiellement remplie**	4.3		#67
53	**Eau coulant dans baignoire remplie**	2.12		#68

n	Label	Durée	Ident	*Ind*
54	**Gouttes dans une baignoire**	3.71		#69
55	**Gouttes dans un évier en métal**	4.89		#70
56	**Eau coulant dans un lavabo**	3.68		#71
57	**Vider un lavabo**	4.84		#72
58	**Jet de douche**	4.2		#73
59	**Remplir une baignoire**	4.51		#74
60	**Filet d'eau dans un évier**	2.69		#75
61	**Remplir un arrosoir**	4.45		#76
62	**Filet d'eau dans un évier**	2.93		#77
63	**Tirer la chasse d'eau**	3.87		#78
	Goutte d'eau sur une plaque chaude	1.34	**X**	#79
64	**Remplir une carafe**	4.95		#80
	Ouvrir une cannette	1.08	**X**	#81

E.2 Fichiers utilisés

Le tableau suivant présente l'origine des sons utilisés dans les expériences du chapitre 5. L'indice *Ind* est l'indice d'origine utilisé avant l'expérience de catégorisation libre.

Ind	Fichier utilisé
# 1	Audio Pro Sound Effects Library/AP11 Household/37 Water Sink.wav
# 2	Audio Pro Sound Effects Library/AP11 Household/41 Water Sink.wav
# 3	Audio Pro Sound Effects Library/AP11 Household/42 Water Sink.wav
# 4	Audio Pro Sound Effects Library/AP11 Household/52 Bathroom Shower.wav
# 5	Audio Pro Sound Effects Library/AP11 Household/53 Bathroom Shower.wav
# 6	Audio Pro Sound Effects Library/AP12 Household, Clocks, Tools/56 Water Bucket.wav
# 7	Auditory Lab/Liquid/Drip audio1/Single drip in Empty Styrofoam /singledrip6 1.wav
# 8	Auditory Lab/Liquid/Drip audio2/Watering Can to Filled Styrofoam Cup/drips(...)2.wav
# 9	Auditory Lab/Liquid/Pour audio1/corn starch solution/pouringgoo2.wav
#10	Auditory Lab/Liquid/Pour audio2/into big plastic tub/pouring into antiEcho bigcontainer4.wav
#11	Auditory Lab/Liquid/Pour audio3/watering can into empty plastic/wateringcan(...)1.wav
#12	Auditory Lab/Liquid/Pour audio3/watering can into empty vase/wateringcanintoemptyredvase1.wav
#13	Auditory Lab/Liquid/Slosh audio1/in 2 liter plastic bottle/sloshingin2liter@120bpm 1.wav
#14	Auditory Lab/Liquid/Slosh audio2/in glass vase/sloshinginredvase3.wav
#15	Auditory Lab/Liquid/Slosh audio2/in milk jug/sloshinginmilkjug@60bmp 1.wav
#16	Auditory Lab/Liquid/Slosh audio3/Wood balls in milk jug/sloshingwoodenballsinmilkjug 2.wav
#17	Auditory Lab/Liquid/Slosh audio3/in styrofoam cup/sloshinginstyrocup9.wav
#18	Auditory Lab/Liquid/Splash audio1/Cardboard/splashingcardboard5.wav
#19	Auditory Lab/Liquid/Splash audio1/fingers in cornstarch solution/splashingfingers ingoo4.wav
#20	Auditory Lab/Liquid/Splash audio1/Metal/splashingmetal2.wav
#21	Auditory Lab/Liquid/Splash audio1/Metal/splashingmetal8.wav
#22	Auditory Lab/Liquid/Splash audio1/fingers in water/splashingfingers inwater3.wav
#23	Auditory Lab/Liquid/Splash audio2/Rigid plastic/splashingplastic3.wav
#24	Auditory Lab/Liquid/Squeeze audio/sponge over water/squeezingspongeoverfulltupperware2.wav
#25	Auditory Lab/Liquid/Squeeze audio/sponge/squeezingsponge2.wav
#26	BBC SFX Library/BBC03 Household/28 Teapot Being Filled With Boiling.wav
#27	BBC SFX Library/BBC03 Household/30 Frying An Egg.wav
#28	BBC SFX Library/BBC03 Household/35 Toilet Flushed.wav
#29	BBC SFX Library/BBC03 Household/39 Bath, Someone Having A Bath 1.wav
#30	BBC SFX Library/BBC03 Household/39 Bath, Someone Having A Bath 2.wav
#31	Soundscan/event classification/liquid/Bubbles/15 Bubble Pop X5 1.wav
#32	Soundscan/event classification/liquid/Bubbles/16 Bubble Various X20 11.wav
#33	Soundscan/event classification/liquid/Bubbles/bubbles02.wav

Ind	Fichier utilisé
#34	Soundscan/event classification/liquid/Fizzing/tocut/bottle rattle fizz.wav
#35	Soundscan/event classification/liquid/Mud Glue/86 Hit Blood Splat X2 1.wav
#36	Soundscan/event classification/liquid/Mud Glue/87 Hit Sloshy Hit X3 2.wav
#37	Soundscan/event classification/liquid/flow/Water Small amount /bottle fill glass4.wav
#38	Blue Box/CD 6/01 water fx/4 3 water dripping.wav
#39	Blue Box/CD 6/01 water fx/4 6 water dripping.wav
#40	Blue Box/CD 6/01 water fx/6 1 water filling.wav
#41	Blue Box/CD 6/01 water fx/9 1 water tap flowing.wav
#42	Hollywood Edge/Hollywood edge/PE 04 Water/08 Drips 7.wav
#43	Hollywood Edge/Hollywood edge/PE 04 Water/14 Single drip into liquid 3.wav
#44	Hollywood Edge/Hollywood edge/PE 04 Water/25 Water plops in metal bucket.wav
#45	Hollywood Edge/Hollywood edge/PE 04 Water/48 Water dripping.wav
#46	Hollywood Edge/Hollywood edge/PE 05 RainThunder Bubbles/05 Rain, roof.wav
#47	Hollywood Edge/Hollywood edge/PE 05 RainThunder Bubbles/24 Bubbling liquid in cauldron.wav
#48	Hollywood Edge/Hollywood edge/PE 16 House Hold/11 Ice Dropping Into Glass Of Water.wav
#49	Hollywood Edge/Hollywood edge/PE 16 House Hold/18 Soda Pouring Into Empty Glass 2.wav
#50	Hollywood Edge/Hollywood edge/PE 16 House Hold/25 Liquid Pouring 3
#51	Hollywood Edge/Hollywood edge/PE 16 House Hold/67 Multiple Water Cooler Glubs.wav
#52	Hollywood Edge/Hollywood edge/PE 16 House Hold/79 Water Going Down Drain 2.wav
#53	Hollywood Edge/Hollywood edge/PE 16 House Hold/84 Shower Running 5.wav
#54	Hollywood Edge/Hollywood edge/PE 16 House Hold/87 Turn On Faucet 1.wav
#55	Hollywood Edge/Hollywood edge/PE 16 House Hold/89 French Fries.wav
#56	Hollywood Edge/Hollywood edge/PE 16 House Hold/90 Hamburgers Fry.wav
#57	Hollywood Edge/Hollywood edge/PE 25 House Hold/31 Boiling Pot.wav
#58	Hollywood Edge/Hollywood edge/PE 25 House Hold/89 Sink.wav
#59	Hollywood Edge/Hollywood edge/PE 25 House Hold/91 Shower.wav
#60	Hollywood Edge/index/CD 25/87 3 Sink Fill.wav
#61	International Sound Effects Library/IN 07 Domestic/51 Audio Track.wav
#62	SoundIdeas/6020/si 20 14 1.wav
#63	SoundIdeas/6020/si 20 15 1.wav
#64	SoundIdeas/6020/si 20 23 1.wav
#65	SoundIdeas/6020/si 20 66 1.wav
#66	SoundIdeas/6021/si 21 66 1.wav
#67	SoundIdeas/6021/si 21 67 1.wav
#68	SoundIdeas/6021/si 21 67 2.wav
#69	SoundIdeas/6021/si 21 69 1.wav
#70	SoundIdeas/6021/si 21 70 1.wav
#71	SoundIdeas/6021/si 21 76 3.wav
#72	SoundIdeas/6021/si 21 78 1.wav
#73	SoundIdeas/6021/si 21 79 1.wav
#74	SoundIdeas/6021/si 21 88 1.wav
#75	Soundscan/19 20 21 SINK FLOW/SINK FLOW 03.wav
#76	Soundscan/22 FILLING POURING/WATER FILLING.wav
#77	Soundscan/23 WATER FLOW/WATER FLOW 3.wav
#78	Soundscan/26 FLUSH/FLUSH 2.wav
#79	Soundscan/27 WATER ON HOT PLATE/WATER ON HOT PLATE02.wav
#80	Soundscan/64 WINE/BOTTLE OF WINE 4.wav
#81	Soundscan/70 MISC BOTTLES AND BOXES/FIZZY BOTTLE OPEN 2.wav

Annexe F

Consignes des expériences

F.1 Expérience préparatoire 1 : Égalisation écologique

Vous allez entendre des sons du quotidien enregistrés dans différents lieux d'habitation. Ces sons ont été enregistrés à des niveaux sonores différents, qui ne correspondent pas forcément à leur vrai niveau dans un environnement réaliste.

Votre tâche est d'ajuster le niveau de ces sons pour qu'il corresponde au mieux au niveau que vous auriez perçu si vous étiez placé dans un contexte naturel d'écoute au sein de l'habitation et positionné à distance de bras de la source sonore.

Durant cette expérience, chaque son sera décrit par une courte phrase, de manière à ce que vous puissiez identifier sa source sonore. Dans un premier temps, vous allez entendre une séquence qui comporte l'ensemble des sons de façon à vous familiariser avec ceux-ci.

Dans un deuxième temps, vous allez entendre des paires de sons composées d'un son de référence et du son à ajuster en niveau. Vous allez ajuster le niveau du second son en le comparant au son de référence, pour que la paire corresponde au mieux à ce que vous auriez perçu si vous étiez placé dans un contexte naturel d'écoute au sein de l'habitation.

Pour ajuster ce deuxième son, vous allez devoir déplacer le curseur qui modifie son niveau sonore. A chaque déplacement du curseur, la paire de son (son de référence / son à ajuster) sera rejouée.

F.2 Expérience préparatoire 2 : Identification

Contexte de l'expérience

Vous allez entendre un ensemble de sons qui ont été enregistrés dans différents lieux d'habitation. Ces sons font intervenir un élément liquide. Ils sont joués à des niveaux sonores différents, comme vous les entendriez dans un contexte naturel d'écoute au sein d'une habitation.

L'expérience

Nous vous demandons si vous parvenez à vous représenter la cause du son. Vous disposez pour cela de cinq étiquettes disposées sur l'écran, de gauche à droite :

–
– « Je ne sais pas du tout».
– « Je ne suis vraiment pas certain(e)»

- « J'hésite entre plusieurs type de causes»
- « Je suis presque certain »
- « Je me représente parfaitement la cause»

Pour chaque son, vous devrez cliquer sur une de ces cinq étiquettes.

Que faut-il comprendre par « se représenter la cause d'un son ? »

Lorsque nous vous demandons de vous représenter la cause du son, nous vous demandons de vous concentrer sur l'événement physique causant le son. Il peut par exemple s'agir de vous représenter qu'un récipient est rempli. Il ne s'agit donc pas de vous représenter toute la situation ou le contexte qui ont pu causer ce son (par exemple, il ne s'agit pas d'identifier qu'un verre est rempli d'eau dans la cuisine).

Comment utiliser l'échelle ?

Pour vous aider à utiliser l'échelle, nous vous présentons deux sons, ainsi que les résultats d'une expérience au cours de laquelle les participants devaient identifier la cause du son. Contrairement à l'expérience que vous allez effectuer, les participants devaient décrire verbalement les sons.

Le son *bien identifié* a été décrit de manière unanime par tous les participants comme : *un robinet est ouvert.* Si vous aussi, vous identifiez parfaitement une cause unique pour ce son, vous devez utiliser la réponse : « J'identifie parfaitement la cause du son »

Le son *mal identifié* a été décrit comme : *Froisser un plastique, Cuire un steak, Essorer une éponge.*

Dans ce cas, les participants identifient plusieurs causes probables. Si vous aussi vous identifiez plusieurs causes alternatives et probables, vous devez alors utiliser la réponse : « J'hésite entre plusieurs causes possibles».

Si par contre, comme certains participants, vous ne parvenez pas à identifier précisément la cause du son, vous devez utiliser la réponse : « Je ne parviens pas à déterminer la cause du son ».

Déroulement de la session

La session est composée de quatre étapes.

- La première étape consiste à écouter tous les sons de l'expérience.
- La seconde étape consiste à vous apprendre à utiliser l'échelle de réponses. Pour cela, nous vous présenterons deux sons correspondant aux degrés extrêmes d'identification de l'échelle.
- La troisième étape consiste à vous familiariser avec l'interface informatique.
- La quatrième étape est l'expérience principale.

Remarques

- Il n'y a pas de « bonne réponse ». N'essayez pas, par exemple, d'équilibrer la proportion de réponses. Si vous identifiez parfaitement tous les sons, vous pouvez très bien répondre «J'identifie parfaitement la cause» de manière systématique. Si vous n'identifiez aucun son, rien ne vous empêche de répondre systématiquement « Je ne sais pas du tout».
- Vous devez essayer de conserver un critère cohérent tout au long de la séance.

F.3 Catégorisation libre

Les consignes des deux parties de l'expérience sont présentées séparément. La consigne de la deuxième partie est présentée après que le participant ait effectué la première partie de l'expérience.

Première Partie

Contexte de l'expérience

Vous allez entendre un ensemble de sons qui ont été enregistrés dans différents lieux d'habitation. Ces sons font intervenir un élément liquide. Ils sont joués à des niveaux sonores différents, comme vous les entendriez dans un contexte naturel d'écoute au sein d'une habitation.

But

Votre but est de former autant de classes que vous le souhaitez, avec autant de sons que vous voulez à l'intérieur de chaque classe.

Les sons doivent être classés en fonction de l'évènement physique qui a causé le son. Par exemple pour les sons : « pétard », « ballon percé », « pneu de voiture qui éclate », l'événement physique qui a produit le son pourrait être explosion.

Vous devez commencer l'expérience par une écoute de tous les sons.

Déroulement de l'expérience : 1ère Partie

- Après une courte phase de familiarisation avec l'interface, vous aurez à former des classes à partir de 64 sons disposés aléatoirement sur l'écran de l'ordinateur.
- Les sons sont matérialisés sous la forme de carrés rouges.
- Lorsque vous cliquez 2 fois sur un carré rouge, vous jouez le son.
- Vous devez écouter tous les sons avant de commencer à les déplacer.
- Lorsque vous cliquez sur un carré rouge et déplacez la souris, vous déplacez le carré rouge sur l'écran. Vous devez déplacer les carrés pour former vos groupes de sons.
- Lorsque votre classification est terminée, vous pouvez sortir de la cabine et demander à l'expérimentateur de venir enregistrer vos résultats.

Remarques :

- Nous vous demandons de ne pas passer plus de 35 min à réaliser cette classification.
- Les sons sont joués à des niveaux sonores différents, comme vous les entendriez dans un contexte naturel d'écoute au sein d'une habitation. Vous ne devez pas essayer de modifier les réglages de la diffusion sonore.

Deuxième Partie

But

Le but est de caractériser précisément chaque classe de sons que vous avez formé dans la première partie. Pour chaque groupe formé, vous allez devoir sélectionner l'élément du groupe le plus représentatif, puis décrire l'ensemble du groupe en quelques lignes.

Déroulement de l'expérience : 2ème Partie

Les sons sont maintenant représentés par des ronds gris. Pour écouter les sons, veuillez cocher la case « Play sound when flying over a button ». Le son sera alors joué quand vous le survolerez avec la souris. Vous pouvez décocher la case « Play sound when flying over a button » quand vous ne souhaitez plus entendre les sons lors du survol avec la souris.

Vous devez effectuer l'ensemble des tâches suivantes de manière itérative :

1. Appuyez sur le bouton « Start Class Definition » en bas à gauche de la fenêtre pour commencer à définir une classe. Cette action ouvre une petite fenêtre sur la droite intitulée « Classes & Comments Definition ».
2. Vous devez ensuite cliquer sur chaque son de la classe pour le sélectionner. Le son sélectionné devient un carré bleu. Vous pouvez désélectionner un son en cliquant dessus une nouvelle fois.
3. Vous devez choisir le son le plus représentatif de cette classe. Pour cela, utilisez le menu déroulant sur la droite « the most typical of this class is » et choisissez le label correspondant au son le plus représentatif de cette classe.
4. Lorsque le son le plus représentatif aura été sélectionné, vous devez décrire votre classe par une brève description.

Lorsque toutes ces tâches ont été accomplies pour un groupe donné, vous pouvez cliquer sur ok dans la petite fenêtre de droite. Vous pouvez alors définir un nouveau groupe de sons en réitérant l'ensemble des 4 tâches ci-dessus.

Le processus s'arrête lorsque tous les groupes ont été définis. Lorsque votre classification est terminée, vous pouvez sortir de la cabine et demander à l'expérimentateur de venir enregistrer vos résultats.

Remarque :

Nous vous demandons de ne pas passer plus de 20 minutes à réaliser cette classification.

Annexe G

Résultats de l'expérience d'égalisation écologique

G.1 Présentation

Les résultats de l'expérience d'égalisation écologique sont présentés dans la figure suivante sous la forme de boîtes à moustaches. Le son 1 est le son de référence qui n'a pas été jugé par les participants. Nous observons pour chaque autre son les valeurs choisies par les participants pour ajuster le volume sonore. Ces valeurs correspondent à une multiplication du signal. Ainsi la valeur 1 correspond au signal identique.

Nous pouvons observer les quartiles (trait continu bleu), le minimum et le maximum (trait pointillé), la médiane (trait rouge) et les valeurs aberrantes (croix rouge) des choix des participants.

G.2 Analyse

Nous pouvons observer une différence de variance des réponses entre les sons. Par exemple, le son 65 (*agiter un verre avec des glaçons*) présente très peu de variance, à part un participant dont la réponse a été jugée *aberrante*. Nous pouvons supposer que la plupart des participants se représentent bien le volume sonore de cette action. Le son 35 également, *presser une matière visqueuse*, présente très peu de variance.

Par contre, d'autres sons, comme les sons de friture (27, 56, 57) présentent une variance très élevée. Les participants imaginent le volume ces sons de manière différente.

Au final les sons ont été modifiés par la valeur médiane.

01 Remplir un lavabo
02 Vider un lavabo
03 Gouttes tombant dans un évier
04 Agiter un pommeau de douche
05 Jet de douche
06 Robinet grand ouvert
07 Gouttes dans un gobelet en plastique
08 Remplir un gobelet en plastique
09 Verser de la peinture dans un pot
10 Verser dans un bac en plastique
11 Verser dans un récipient vide en plastique
12 Verser dans un vase vide
13 Secouer une bouteille en plastique remplie
14 Secouer un vase contenant de l'eau
15 Secouer un bidon en plastique rempli
16 Secouer un bidon contenant du liquide et des solides
17 Secouer un gobelet en plastique
18 Eclaboussures sur du carton
19 Mélanger de la peinture
20 Eclaboussures sur une plaque de métal
21 Verser sur une plaque de métal
22 Agiter de l'eau avec ses doigts
23 Eclaboussures sur du plastique
24 Presser une éponge au dessus de l'évier
25 Presser une éponge
26 Remplir une théière
27 Faire frire un oeuf
28 Tirer la chasse d'eau
29 Bouger dans un bain
30 Objet tombant dans le bain
31 Bulles dans l'eau
32 Ébullition dans une grande casserole
33 Souffler avec une paille à la surface de l'eau
34 Secouer une bouteille en plastique
35 Presser une matière visqueuse
36 Impact de matière visqueuse
37 Vider une bouteille
38 Gouttes dans un évier
39 Flot d'eau dans un grand évier
40 Remplir un bidon en plastique
41 Remplir un évier
42 Gouttes tombant sur des feuilles
43 Goutte tombant dans l'eau
44 Agiter un récipient en métal à moitié rempli
45 Essorer des habits mouillés
46 Pluie tombant sur le sol depuis le toit
47 Ébullition dans un chaudron
48 Mettre un glaçon dans une boisson
49 Pétillement après ouverture d'une cannette
50 Remplir un verre contenant des glaçons
51 Agiter un récipient en métal à moitié rempli
52 Vider un lavabo
53 Jet de douche dans une baignoire
54 Ouvrir un robinet
55 Huile de friture bouillante
56 Steak frit sur un grill
57 Eau bouillante dans une casserole
58 Vider un évier
59 Jet de douche avec le bruit d'évacuation
60 Remplir un évier
61 Verser du ketchup
62 Pulvériser de l'eau
63 Pulvériser dans un verre
64 Verser de la bière dans un verre
65 Agiter un verre contenant des glaçons
66 Remplir une baignoire
67 Eau coulant dans baignoire partiellement remplie
68 Eau coulant dans baignoire remplie
69 Gouttes dans une baignoire
70 Gouttes dans un évier en métal
71 Eau coulant dans un lavabo
72 Vider un lavabo
73 Jet de douche
74 Remplir une baignoire
75 Filet d'eau dans un évier
76 Remplir un arrosoir
77 Filet d'eau dans un évier
78 Tirer la chasse d'eau
79 Goutte d'eau sur une plaque chaude
80 Remplir une carafe
81 Ouvrir une cannette

Annexe H

Verbalisations

Les verbalisations des participants issues de l'expérience de catégorisation sont présentées dans les deux tableaux suivants. Dans ces tableaux, nous avons omis les termes communs à l'ensemble des classes : *eau, bruit, son, liquide*. Pour chaque classe de sons, nous présentons les termes principaux ainsi que leurs occurrences (*Occ*). Ces termes sont classés par ordre de spécificité (*Spec*) décroissante.

Classe A			**Classe B**			**Classe C**		
Label	**Occ**	**Spec**	**Label**	**Occ**	**Spec**	**Label**	**Occ**	**Spec**
chasse d'eau	75	*57.9*	douche:doucher	273	*24.7*	goutter:goutte:goutte à goutte	145	*18.5*
aspirer:aspiration	41	*37.4*	jet	108	*18.4*	objet	32	*16.0*
evacuer:evacuation	35	*25.0*	fort	97	*15.3*	essorer:essorage	33	*14.1*
syphon	20	*15.1*	pleuvoir:pluie	89	*10.2*	main	30	*12.7*
tirer	13	*9.4*	couler	144	*9.8*	secouer	29	*12.2*
vider	26	*7.8*	continu	68	*8.3*	mouvement	48	*11.5*
style	6	*6.7*	presser:pression	93	*7.9*	remuer	24	*9.8*
wc	11	*6.5*	laver	48	*6.6*	linge	15	*7.5*
tuyau	9	*5.9*	remplir:remplissage	127	*6.6*	mal fermé	16	*6.9*
canalisation	6	*5.8*	haut	31	*5.6*	éponge	16	*6.1*

Classe B_1			**Classe** B_2			**Classe** B_3		
Label	**Occ**	**Spec**	**Label**	**Occ**	**Spec**	**Label**	**Occ**	**Spec**
baignoire	31	6.2	uriner	24	14.2	verre	66	30.8
ouvert:ouverture:ouvrir	21	5.3	filet	32	8.3	verser:versement	60	20.2
lavabo	20	5.3	surface	46	8.0	récipient	48	16.3
robinet	47	5.3	correspondre	20	6.9	remplir:remplissage	45	10.8
degré	6	4.1	sur	54	5.9	carafe	16	10.4
vasque	5	3.8	faible	14	5.6	dans	107	7.4
continu	21	3.7	directement	7	5.2	boire	11	6.4
remplir:remplissage	36	3.7	légerement	7	5.2	boisson	12	6.3
haut	12	3.6	couler	45	4.9	servir	12	6.1
couler	38	3.6	fil	7	4.6	contenir	16	5.7
Classe B_4			**Classe** B_5			**Classe** B_6		
Label	**Occ**	**Spec**	**Label**	**Occ**	**Spec**	**Label**	**Occ**	**Spec**
doucher	178	60.5	puissant	4	3.3	ebullition	20	18.4
fort	53	16.0	jet	10	3.1	bulle	19	10.5
pluie	46	12.3	nettoyer	2	2.4	bouillir	16	7.5
presser:pression	47	10.5	animal	2	2.2	porter	6	7.4
pommeau	23	10.1	grogner	2	2.2	boue	8	6.2
jet	47	10.0	vapeur	2	2.2	chauffer	7	4.6
prendre	15	8.5	bassine	2	2.0	paille	6	4.1
laver	27	8.4	fort	7	1.7	conséquence	4	3.8
cabine	11	6.9	remplir:remplissage	9	1.6	facteur	4	3.8
haut	18	6.5	aspirer	4	1.4	naturel	4	3.8

Annexe I

Expérience préliminaire d'identification d'activité

Cette annexe décrit une expérience préliminaire effectuée dans l'objectif d'évaluer les difficultés à reconnaître une activité dans une liste donnée, à partir de l'audio seulement.

Protocole de test

Dans un premier temps, les participants doivent identifier l'activité à partir d'un extrait audio. Ils peuvent ainsi choisir l'une des réponses suivantes :

- se laver les mains,
- se laver les dents,
- préparer à manger/boire,
- jardiner,
- je ne sais pas.

Dans un deuxième temps, les participants indiquent les raisons qui ont permis de choisir cette activité. Les participants peuvent ainsi choisir des causes dans la liste suivante (établie arbitrairement, ce test ayant été effectué avant l'établissement des catégories perceptives) :

- Gouttes
- Jet / flots,
- Choc / impacts,
- Frottement,
- Versement,
- Écoulement,
- Éclaboussure / splash / plouf,
- Remplir,
- Autres.

Les participants ont également la possibilité de décrire librement d'autres raisons.

Dans un troisième temps les participants ont accès à la vidéo de l'extrait, et doivent à nouveau identifier l'activité dans la liste présentée dans la première partie.

Résultats

Dans ce test, 19 extraits d'activités issus du corpus IMMED ont été présentés. Onze participants de notre équipe de recherche ont effectué cette expérience.

Au final, si certaines activités peuvent être très bien reconnues par l'audio, comme par exemple *se laver les dents*, d'autres activités, comme *préparer à manger* sont plus difficilement identifiables. En moyenne, la similarité entre l'identification de l'activité par l'audio seul, et de son identification par l'audio et la vidéo est de 63%.

Ce score traduit des résultats d'identification assez faibles. De plus, dans ce test, les participants n'utilisent pas uniquement les sons d'eau, mais aussi les chocs (produits par exemple par la vaisselle) et la parole. Près de 10% des réponses ont ainsi été influencées par la compréhension de la parole.

Bibliographie

[ADD+88] Jean-Marie Aran, A. Dancer, J.M. Dolmazon, R. Pujol, and P. Tran Ba Huy. *Physiologie de la cochlée*. INSERM/SFA Série Audition, 1988.

[AFF+11] V. Akkermans, F. Font, J. Funollet, B. De Jong, G. Roma, S. Togias, and X. Serra. Freesound 2: An improved platform for sharing audio clips. In *International Society for Music Information Retrieval Conference, ISMIR, Late-breaking Demo Session*, 2011.

[Alz13] La Fondation Plan Alzheimer. Le plan alzheimer 2008-2012. `http://www.fondation-alzheimer.org/content/le-plan-alzheimer-2008-2012`, consulté le 8 avril 2013.

[AMSMH08] M. Al Masum Shaikh, M. Khademul Islam Molla, and Keikichi Hirose. Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense. In *Proceedings of the 11th International Conference on Computer and Information Technology, ICCIT*, pages 294–299. IEEE, 2008.

[AO88] Régine André-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 36(1):29–40, 1988.

[AP10] Salvatore M. Aglioti and Mariella Pazzaglia. Representing actions through their sound. *Experimental brain research*, 206(2):141–151, 2010.

[AVCC07] Hervé Abdi, Dominique Valentin, Sylvie Chollet, and Christelle Chrea. Analyzing assessors and products in sorting tasks: Distatis, theory and applications. *Food quality and preference*, 18(4):627–640, 2007.

[AVRP13] Frédéric Aman, Michel Vacher, Solange Rossato, and François Portet. In-home detection of distress calls: the case of aged users. In *Proceedings of the the Annual Conference of International Speech Communication Association, INTERSPEECH.*, volume 500, page 202, 2013.

[Bal93] James A. Ballas. Common factors in the identification of an assortment of brief everyday sounds. *Journal of experimental psychology: human perception and performance*, 19(2):250, 1993.

[Bar83] Lawrence W. Barsalou. Ad hoc categories. *Memory & cognition*, 11(3):211–227, 1983.

[BCEG07] Sylvain Bonhomme, Eric Campo, Daniel Estève, and Joelle Guennec. An extended prosafe platform for elderly monitoring at home. In *Proceedings of the 29th Annual International Conference of the Engineering in Medicine and Biology Society, EMBS*, pages 4056–4059. IEEE, 2007.

[BDVA+08] Marinus M. Boone, Diemer De Vries, Tjeerd C. Andringa, Anton Schlesinger, Jasper Van Dorp Schuitman, Bea Valkenier, and Hedde Van De Vooren. Modelling of the cochlea response as a versatile tool for acoustic signal processing. *Journal of the Acoustical Society of America*, 123(5):3722, 2008.

[Ber09] Nancy Bertin. *Les factorisations en matrices non-négatives. Approches contraintes et probabilistes, application à la transcription automatique de musique polyphonique.* PhD thesis, Télécom ParisTech, 2009.

[Bis95] Christopher M. Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

[Bjö86] E. A. Björk. Laboratory annoyance and skin conductance responses to some natural sounds. *Journal of Sound and Vibration*, 109(2):339–345, 1986.

[BM91] James A. Ballas and Timothy Mullins. Effects of context on the identification of everyday sounds. *Human performance*, 4(3):199–219, 1991.

[Bra21] Sir W. H. Bragg. *The World of Sound*. Bell, London, 1921.

[Bre93] Leo Breiman. *Classification and regression trees*. CRC press, 1993.

[Bre94] Albert S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[BS95] Anthony J. Bell and Terrence J. Sejnowski. Blind separation and blind deconvolution: an information-theoretic approach. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 5, pages 3415–3418. IEEE, 1995.

[Bur98] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[BUS+11] Joydip Barman, Gitendra Uswatte, Nilanjan Sarkar, Touraj Ghaffari, and Brad Sokal. Sensor-enabled rfid system for monitoring arm activity in daily life. In *Proceedings of the Annual International Conference of the Engineering in Medicine and Biology Society, EMBC*, pages 5219–5223. IEEE, 2011.

[BW01] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[BWK+10] Rolf Bardeli, D. Wolff, Frank Kurth, M. Koch, K. H. Tauchert, and K. H. Frommolt. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31(12):1524–1534, 2010.

[CBdL99] José Luis Carles, Isabel López Barrio, and José Vicente de Lucio. Sound influence on landscape values. *Landscape and urban planning*, 43(4):191–200, 1999.

[CEEC08] Marie Chan, Daniel Estève, Christophe Escriba, and Eric Campo. A review of smart homes—Present state and future challenges. *Computer methods and programs in biomedicine*, 91:55–81, 2008.

[CER05] Chloé Clavel, Thibaut Ehrette, and Gaël Richard. Events detection for an audio-based surveillance system. In *Proceedings of the International Conference on Multimedia and Expo, ICME*, pages 1306–1309. IEEE, 2005.

[Che57] Colin Cherry. *On human communication; a review, a survey, and a criticism*. The Technology Press of MIT, 1957.

[CHEI94] B. G. Celler, T. Hesketh, W. Earnshaw, and E. Ilsar. An instrumentation system for the remote monitoring of changes in functional health status of the elderly at home. In *Proceedings of the Annual International Conference of the Engineering in Medicine and Biology Society, EMBC*, volume 2, pages 908–909. IEEE, 1994.

[CHRC95] M. Chan, C. Hariton, P. Ringeard, and E. Campo. Smart house automation system for the elderly and the disabled. In *Proceedings of the International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, volume 2, pages 1586–1589. IEEE, 1995.

[CI01] Eric Castelli and Dan Istrate. Everyday life sounds and speech analysis for a medical telemonitoring system. In *Proceedings of the 7th European Conference on Speech Communication and Technology, EUROSPEECH*, pages 2417–2420, 2001.

[CKZ+05] Jianfeng Chen, Alvin Harvey Kam, Jianmin Zhang, Ning Liu, and Louis Shue. Bathroom activity monitoring based on sound. In *Pervasive Computing*, pages 47–61. Springer, 2005.

[CLC+10] T. Campbell, E. Larson, G. Cohn, J. Froehlich, R. Alcaide, and S.N. Patel. WATTR: a method for self-powered wireless sensing of water activity in the home. In *Proceedings of the 2010 International conference on Ubiquitous computing*, pages 169–172. ACM, 2010.

[CLS10] Chris Cannam, Christian Landone, and Mark Sandler. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the international conference on Multimedia*, pages 1467–1468. ACM, 2010.

[CNK09] Selina Chu, Shrikanth Narayanan, and C-CJ Kuo. Environmental sound recognition with time–frequency audio features. *Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009.

[CP00] Patrick A. Cabe and John B. Pittenger. Human sensitivity to acoustic information from vessel filling. *Journal of experimental psychology: human perception and performance*, 26(1):313, 2000.

[DBB52] K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24:637, 1952.

[dBHU+08] Eling de Bruin, Antonia Hartmann, Daniel Uebelhart, Kurt Murer, and Wiebren Zijlstra. Wearable systems for monitoring mobility-related activities in older people: a systematic review. *Clinical rehabilitation*, 22(10-11):878–895, 2008.

[DCK02] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111:1917, 2002.

[DLR77] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[DNM+13] Daniel Diep, Hervé Nonon, Isabelle Marc, Jonathan Delhom, Frédéric Roure, et al. Acoustic counting and monitoring of shad fish populations. In *International AmiBio Workshop: Recent Progress in Computational Bioacoustics for Assessing Biodiversity*, 2013.

[Dov11] Vladislavs Dovgalecs. *Indoor location estimation using a wearable camera with application to the monitoring of persons at home.* PhD thesis, Université Sciences et Technologies-Bordeaux I, 2011.

[Dub93] Danièle Dubois. *Sémantique et cognition - Catégories, prototypes, typicalité*. CNRS, Paris, 1993.

[Dub00] Danièle Dubois. Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive Science Quaterly*, 1:35–68, 2000.

[Ell96] Daniel Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, 1996.

[Esc73] Yves Escoufier. Le traitement des variables vectorielles. *Biometrics*, pages 751–760, 1973.

[FAH06] J. Fogarty, C. Au, and S.E. Hudson. Sensing from the basement: a feasibility study of unobtrusive and low-cost home activity recognition. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, 2006.

[FAP08] Homa Foroughi, Baharak Shakeri Aski, and Hamidreza Pourreza. Intelligent video surveillance for monitoring fall detection of elderly in home environments. In *Proceedings of the 11th International Conference on Computer and Information Technology, ICCIT*, pages 219–224. IEEE, 2008.

[FLC+09] J. Froehlich, E. Larson, T. Campbell, C. Haggerty, J. Fogarty, and S.N. Patel. Hydrosense: infrastructure-mediated single-point sensing of whole-home water activity. In *Proceedings of the 2009 ACM International Conference on Ubiquitous Computing*, pages 235–244. ACM, 2009.

[Fra59] G. J. Franz. Splashes as sources of sound in liquids. *The Journal of the Acoustical Society of America*, 31:1080, 1959.

[Fri13] J. Fricke. Les activités de la vie quotidienne. *J.H. Stone, M. Blouin, International Encyclopedia of Rehabilitation. Available online : http://cirrie.buffalo.edu/encyclopedia/fr/article/37/*, 2013.

[Gai09] Pascal Gaillard. Laissez-nous trier ! TCL-LabX et les tâches de catégorisation libre de sons. *Le sentir et le dire. Concepts et méthodes en psychologie et linguistique cognitive*, 2009.

[Gav93] William W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993.

[GBS+13] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark Plumbley. Detection and classification of acoustic scenes and events. *An IEEE AASP Challenge*, 2013.

[GD99] Gilles Gonon and Claude Depollier. Estimation des paramètres d'un sinus glissant par transformé de fourier fractionnaire. In *17ème Colloque sur le traitement du signal et des images*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 1999.

[GGM+05] Sylvain Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, 2005.

[GGWM11] M.N. Geffen, J. Gervain, J.F. Werker, and M.O. Magnasco. Auditory perception of self-similarity in water sounds. *Frontiers in Integrative Neuroscience*, 5, 2011.

[GHNO02] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical and jazz music databases. In *Proceedings of the*

3rd International Conference on Music Information Retrieval, ISMIR, volume 2, pages 287–288, 2002.

[GKW07] Brian Gygi, Gary R Kidd, and Charles S Watson. Similarity and categorization of environmental sounds. *Perception & psychophysics*, 69(6):839–855, 2007.

[GM06] Bruno L. Giordano and Stephen McAdams. Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, 119:1171, 2006.

[GM07] Ilan Gronau and Shlomo Moran. Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters*, 104(6):205–210, 2007.

[GMM10] Bruno L. Giordano, John McDonnell, and Stephen McAdams. Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds. *Brain and cognition*, 73(1):7–19, 2010.

[Gol94] Robert L. Goldstone. The role of similarity in categorization: Providing a groundwork. *Cognition*, 52(2):125–157, 1994.

[GPAO12] Patrice Guyot, Julien Pinquier, and Régine Andre-Obrecht. Water flow detection from a wearable device with a new feature, the spectral cover. In *Proceedings of the 10th International Workshop on Content-Based Multimedia Indexing, CBMI*. IEEE, 2012.

[GPAO13] Patrice Guyot, Julien Pinquier, and Régine André-Obrecht. Water sound recognition based on physical models. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. IEEE, 2013.

[GPD00] Fabien Gouyon, François Pachet, and Olivier Delerue. On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the COST G-6 conference on Digital Audio Effects, DAFX*. Citeseer, 2000.

[Guy96] Frédérique Guyot. *Etude de la perception sonore en termes de reconnaissance et d'appréciation qualitative: une approche par la catégorisation.* PhD thesis, Université du Maine, 1996.

[GVPA13] Patrice Guyot, Xavier Valero, Julien Pinquier, and Francesc Alías. Two-step detection of water sound events for the diagnostic and monitoring of dementia. In *Proceedings of the International Conference on Multimedia and Expo, ICME*. IEEE, 2013.

[GZL01] Guodong Guo, HongJiang Zhang, and Stan Z Li. Boosting for content-based audio classification and retrieval: An evaluation. In *Proceedings of the International Conference on Multimedia and Expo, ICME*. IEEE, 2001.

[HAAH10] F. Hijaz, N. Afzal, T. Ahmad, and O. Hasan. Survey of fall detection and daily activity monitoring techniques. In *Proceedings of the International Conference on Information and Emerging Technologies, ICIET*, pages 1–6, 2010.

[HB80] James H. Howard and James A. Ballas. Syntactic and semantic factors in the classification of nonspeech transient patterns. *Perception & Psychophysics*, 28(5):431–439, 1980.

[HC12] Qiang Huang and Stephen Cox. Improved audio event detection by use of contextual noise. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 493–496. IEEE, 2012.

[Hel63] Hermann Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press, 1863.

[Her90] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738, 1990.

[Hes83] Wolfgang Hess. *Pitch determination of speech signals: algorithms and devices*. Springer-Verlag, 1983.

[HLDC+04] Pierre Hanna, Nicolas Louis, Myriam Desainte-Catherine, Jenny Benois-Pineau, et al. Audio features for noisy sound segmentation. In *Proceedings of the 3rd International Conference on Music Information Retrieval, ISMIR*, 2004.

[HLM+12] Olivier Houix, Guillaume Lemaitre, Nicolas Misdariis, Patrick Susini, and Isabel Urdapilleta. A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied*, 18(1):52, 2012.

[HMP10] Serge Heiden, Jean-Philippe Magué, and Bénédicte Pincemin. TXM : Une plateforme logicielle open-source pour la textométrie - Conception et développement. In Luca Giuliano Sergio Bolasco, Isabella Chiari, editor, *Proceedings of 10th International Conference on the Statistical Analysis of Textual Data*, volume 2, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto, 2010.

[HMS05] Aki Harma, Martin F McKinney, and Janto Skowronek. Automatic surveillance of the acoustic activity in our living environment. In *Proceedings of the International Conference on Multimedia and Expo, ICME*. IEEE, 2005.

[HMVE11] Toni Heittola, Annamaria Mesaros, Tuomas Virtanen, and Antti Eronen. Sound event detection in multisource environments using source separation. In *Workshop on machine listening in Multisource Environments*, pages 36–40, 2011.

[HWB+06] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. SenseCam: a retrospective memory aid. In *Proceedings of the International conference on Ubiquitous computing*, pages 177–193. Springer Berlin Heidelberg, 2006.

[IBC+08] A. Ibarz, G. Bauer, R. Casas, A. Marco, and P. Lukowicz. Design and evaluation of a sound based water flow measurement system. In *Proceedings of the Third European Conference of Smart Sensing and Context, EuroSSC*, pages 41–54. Springer, 2008.

[ICV+06] Dan Istrate, Eric Castelli, Michel Vacher, Laurent Besacier, and J-F Serignat. Information extraction from sound for medical telemonitoring. *Transactions on Information Technology in Biomedicine*, 10(2):264–274, 2006.

[JD88] Anil Jain and Richard Dubes. *Algorithms for clustering data*. Prentice Hall Englewood Cliffs, 1988.

[Joh88] J.D. Johnston. Transform coding of audio signals using perceptual noise criteria. *Selected Areas in Communications*, 1988.

[Kar11] Svebor Karaman. *Indexation de la vidéo Portée: Application à l'étude épidémiologique des maladies liées à l'âge*. PhD thesis, Université Sciences et Technologies-Bordeaux I, 2011.

[Kat83] Sidney Katz. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*, 1983.

[KBPD+11] Svebor Karaman, Jenny Benois-Pineau, Vladislavs Dovgalecs, Rémi Mégret, Julien Pinquier, Régine André-Obrecht, Yann Gaëstel, and Jean-François Dartigues. Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia. *Multimedia Tools and Applications*, pages 1–29, 2011.

[KBPM+10] Svebor Karaman, Jenny Benois-Pineau, Remi Megret, Vladislavs Dovgalecs, J-F. Dartigues, and Yann Gaestel. Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In *Proceedings of the 20th International Conference on Pattern Recognition, ICPR*, pages 4113–4116. IEEE, 2010.

[KD06] Anssi Klapuri and Manuel Davy. *Signal processing methods for music transcription*. Springer, 2006.

[KG09] K. Kappel and T. Grechenig. Show-me: water consumption at a glance to promote water conservation in the shower. In *Proceedings of the 4th International Conference on Persuasive Technology*, page 26. ACM, 2009.

[KMS06] Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora. *MPEG-7 audio and beyond: Audio content indexing and retrieval*. Wiley, 2006.

[KNA09] Dirkjan Krijnders, Maria E. Niessen, and Tjeerd C. Andringa. Annotating soundscapes. In *Proceedings of the International Congress on Noise Control Engineering, INTERNOISE*, 2009.

[Kof35] Kurt Koffka. *Principles of Gestalt psychology*. Harcourt, Brace New York, 1935.

[KPT00] Andrew J. Kunkler-Peck and M. T. Turvey. Hearing shape. *Journal of Experimental psychology: human perception and performance*, 26(1):279, 2000.

[KS04] Hyoung-Gook Kim and Thomas Sikora. How efficient is mpeg-7 for general sound recognition? In *Audio Engineering Society Conference: 25th International Conference: Metadata for Audio*. Audio Engineering Society, 2004.

[LAC07] Nelly Christina Laydrus, Eliathamby Ambikairajah, and Branko Celler. Automated sound analysis system for home telemonitoring using shifted delta cepstral features. In *Proceedings of the 15th International Conference on Digital Signal Processing*, pages 135–138. IEEE, 2007.

[Laf84] Pierre Lafon. *Dépouillements et statistiques en lexicométrie*, volume 24. Slatkine, 1984.

[LALM05] Christophe Laplanche, Olivier Adam, Maciej Lopatka, and Jean-François Motsch. Male sperm whale acoustic behavior observed from multipaths at a single hydrophone. *The Journal of the Acoustical Society of America*, 118:2677, 2005.

[Lei97] T. G. Leighton. *The acoustic bubble*, volume 10. Academic Press, 1997.

[LH12] Guillaume Lemaitre and Laurie M Heller. Auditory perception of material is fragile while action is strikingly robust. *The Journal of the Acoustical Society of America*, 131:1337, 2012.

[LHMS10] Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and Patrick Susini. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, 16(1):16, 2010.

[LKD10] Richard F. Lyon, Andreas G. Katsiamis, and Emmanuel M. Drakakis. History and future of auditory filter models. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems, ISCAS*, pages 3809–3812. IEEE, 2010.

[LS99] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[LSB94] Ludovic Lebart, André Salem, and Christian Baudelot. *Statistique textuelle*, volume 342. Dunod París, 1994.

[LSDM01] Dongge Li, Ishwar K. Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern recognition letters*, 22(5):533–544, 2001.

[LT07] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[Ltd13] Audio Analytic Ltd. `http://www.audioanalytic.com/`, consulté le 30 août 2013.

[Lut07] Robert A Lutfi. Human sound source identification. In *Auditory perception of sound sources*, pages 13–42. Springer, 2007.

[LZL03] Lie Lu, Hong-Jiang Zhang, and Stan Z Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia systems*, 8(6):482–492, 2003.

[MA09] Ghulam Muhammad and Khaled Alghathbar. Environment recognition from audio using mpeg-7 features. In *4th International Conference on Embedded and Multimedia Computing, EM-Com*, pages 1–6. IEEE, 2009.

[MAYKM13] Adrien Merer, Mitsuko Aramaki, Sølvi Ystad, and Richard Kronland-Martinet. Perceptual characterization of motion evoked by sounds for synthesis control purposes. *ACM Transactions on Applied Perception, TAP*, 10(1):1, 2013.

[MBCH08] Alex Mihailidis, Jennifer N Boger, Tammy Craig, and Jesse Hoey. The coach prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC geriatrics*, 8(1):28, 2008.

[MBG+00] Michael M. Marcell, Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers. Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology*, 22(6):830–864, 2000.

[MDF+84] Guy McKhann, David Drachman, Marshall Folstein, Robert Katzman, Donald Price, and Emanuel M. Stadlan. Clinical diagnosis of alzheimer's disease report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology*, 34(7):939–939, 1984.

[MDW+10] Rémi Mégret, V. Dovgalecs, H. Wannous, S. Karaman, J. Benois-Pineau, E. El Khoury, J. Pinquier, P. Joly, R. André-Obrecht, and Y. Gaëstel. The IMMED project: wearable video monitoring of people with age dementia. In *Proceedings of the International Conference on Multimedia*, pages 1299–1302. ACM, 2010.

[MEF+10] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. YAAFE, an easy to use and efficient audio feature extraction software. In *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, pages 441–446, 2010.

[MHEV10] Annamaria Mesaros, Toni Heittola, Antti Eronen, and Tuomas Virtanen. Acoustic event detection in real life recordings. In *Proceedings of the 18th European Signal Processing Conference*, pages 1267–1271, 2010.

[MIBD09] H. Medjahed, D. Istrate, J. Boudy, and B. Dorizzi. Human activities of daily living recognition using fuzzy logic for elderly home monitoring. In *Proceedings of the International Conference on Fuzzy Systems*, pages 2001–2006. IEEE, 2009.

[Min33] M. Minnaert. On musical air-bubbles and the sounds of running water. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 16(104):235–248, 1933.

[MMTN05] Robert Malkin, Dušan Macho, Andrey Temko, and Climent Nadeu. First evaluation of acoustic event classification systems in CHIL project. In *Proceedings of the Workshop on Hands-free Speech Communication and Microphone Arrays, HSCMA*, 2005.

[Mor90] Edgar Morin. Carrefour des sciences. In *Actes du colloque du Comité national de la Recherche scientifique interdisciplinaire*, 1990.

[MPG12] Brian A. Mazzeo, Anjali N. Patil, and W. Spencer Guthrie. Acoustic impact-echo investigation of concrete delaminations using liquid droplet excitation. *NDT & E International*, 51:41–44, 2012.

[MYH+10] W. Moss, H. Yeh, J.M. Hong, M.C. Lin, and D. Manocha. Sounding liquids: Automatic sound synthesis from fluid simulation. *ACM Transactions on Graphics, TOG*, 29(3):21, 2010.

[Nat03] United Nations. International year of freshwater 2003. `http://www.un.org/events/water/`, 2003.

[NFN+09] N. Noury, A. Fleury, R. Nocua, J. Poujaud, C. Gehin, A. Dittmar, G. Delhomme, J. Demongeot, and E. McAdam. Capteurs pour la télésurveillance médicale. capteurs, algorithmes et réseaux. *IRBM*, 30(3):93–103, 2009.

[NHA+00] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC*, 2000.

[OMCV13] Koray Ozcan, Anvith Katte Mahabalagiri, Mauricio Casares, and Senem Velipasalar. Automatic fall detection and activity classification by a wearable embedded smart camera. In *Proceedings of the International Conference on Multimedia and Expo, ICME*. IEEE, 2013.

[PCST99] John C. Platt, Nello Cristianini, and John Shawe-Taylor. Large margin dags for multiclass classification. In *Neural Information Processing Systems, NIPS*, volume 12, pages 547–553, 1999.

[Pee04] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. *CUIDADO Project*, 2004.

[PFP+04] Matthai Philipose, Kenneth P Fishkin, Mike Perkowitz, Donald J Patterson, Dieter Fox, Henry Kautz, and Dirk Hahnel. Inferring activities from interactions with objects. *Pervasive Computing*, 3(4):50–57, 2004.

[PHA+08] Karine Peres, Catherine Helmer, Helene Amieva, Jean-Marc Orgogozo, Isabelle Rouch, Jean-François Dartigues, and Pascale Barberger-Gateau. Natural history of decline in instrumental activities of daily living performance over the 10 years preceding the clinical diagnosis of dementia: A prospective population-based study. *Journal of the American Geriatrics Society*, 56(1):37–44, 2008.

[Pin04] Julien Pinquier. *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle*. PhD thesis, Université Paul Sabatier-Toulouse III, 2004.

[PKL+12] Julien Pinquier, Svebor Karaman, Laetitia Letoupin, Patrice Guyot, Rémi Mégret, Jenny Benois-Pineau, Yann Gaestel, and Jean-François Dartigues. Strategies for multiple feature fusion with hierarchical hmm: application to activity recognition from wearable audiovisual sensors. In *Proceedings of the 21th International Conference on Pattern Recognition, ICPR*, pages 3192–3195. IEEE, 2012.

[PLSR08] Mihail Popescu, Yun Li, Marjorie Skubic, and Marilyn Rantz. An acoustic fall detector system that uses sound height information to reduce the false alarm rate. In *Proceedings of the 30th Annual International Conference of the Engineering in Medicine and Biology Society, EMBC*, pages 4628–4631. IEEE, 2008.

[PLST09] Ya-Ti Peng, Ching-Yung Lin, Ming-Ting Sun, and Kun-Cheng Tsai. Healthcare audio event classification using hidden markov models and hierarchical hidden markov models. In *Proceedings of the International Conference on Multimedia and Expo, ICME*, pages 1218–1221. IEEE, 2009.

[PNB+07] Lorenzo Piccardi, Basilio Noris, Olivier Barbey, Aude Billard, Giuseppina Schiavone, Flavio Keller, and Claes von Hofsten. Wearcam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children. In *Proceedings of the 16th International Symposium on Robot and Human interactive Communication, RO-MAN*, pages 594–598. IEEE, 2007.

[PPK01] Daniel Pressnitzer, Roy D. Patterson, and Katrin Krumbholz. The lower limit of melodic pitch. *The Journal of the Acoustical Society of America*, 109(5):2074, 2001.

[PR07] François Pachet and Pierre Roy. Exploring billions of audio features. In *Proceedings of the 7th International Workshop on Content-Based Multimedia Indexing, CBMI*, pages 227–235. IEEE, 2007.

[PTK+02] Vesa Peltonen, Juha Tuomi, Anssi Klapuri, Jyri Huopaniemi, and Timo Sorsa. Computational auditory scene recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. IEEE, 2002.

[Qua13] Quaero. `http://www.quaero.org/`, consulté le 15 juin 2013.

[Ram10] Ali Rammal. *Modélisation multi-agent dans un processus de gestion multi acteur, application au maintien à domicile*. PhD thesis, Université Paul Sabatier-Toulouse III, 2010.

[RC04] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

[Rei93] Martin Rein. Phenomena of liquid drop impact on solid and liquid surfaces. *Fluid Dynamics Research*, 12(2):61–93, 1993.

[Rex11] James Rex. Audio event detection and localisation for directing video surveillance-a survey of potential techniques. In *Proceedings of the 4th International Conference on Imaging for Crime Detection and Prevention, ICDP*, pages 1–4. IET, 2011.

[RJK+10a] Gerard Roma, Jordi Janer, Stefan Kersten, Mattia Schirosa, and Perfecto Herrera. Content-based retrieval from unstructured audio databases using an ecological acoustics taxonomy. In *Proceedings of the 16th International Conference on Auditory Display, ICAD*, 2010.

[RJK+10b] Gerard Roma, Jordi Janer, Stefan Kersten, Mattia Schirosa, Perfecto Herrera, and Xavier Serra. Ecological acoustics perspective for content-based retrieval of environmental sounds. *Journal on Audio, Speech, and Music Processing, EURASIP*, 2010:7, 2010.

[RL78] Eleanor Rosch and Barbara B Lloyd. *Cognition and categorization*. Hillsdale, New Jersey, 1978.

[RL08] Mary Vining Radomski and Catherine A Trombly Latham. *Occupational Therapy for Physical Dysfunction, sixth edition*. Lippincott Williams & Wilkins, 2008.

[RM75] Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

[RM99] Brian H. Ross and G. L. Murphy. Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, 38(4):495–553, 1999.

[Roa88] Curtis Roads. Introduction to granular synthesis. *Computer Music Journal*, 12(2):11–13, 1988.

[Ros73] Eleanor H Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973.

[RP13] Zafar Rafii and Bryan Pardo. Online repet-sim for real-time speech enhancement. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. IEEE, 2013.

[SB03] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180. IEEE, 2003.

[SBB+07] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan. The CLEAR 2006 evaluation. In *Multimodal Technologies for Perception of Humans*, pages 1–44. Springer, 2007.

[SC67] Pierre Schaeffer and Michel Chion. *La musique concrète*. Presses universitaires de France, 1967.

[SCB+06] Cliodhna Ní Scanaill, Sheila Carew, Pierre Barralon, Norbert Noury, Declan Lyons, and Gerard M. Lyons. A review of approaches to mobility telemonitoring of the elderly in their living environment. *Annals of Biomedical Engineering*, 34(4):547–563, 2006.

[Sch77] R. Murray Schafer. *The soundscape*. J.-C. Lattès, 1977.

[Ser96] United States. Public Health Service. *Physical activity and health: a report of the Surgeon General*. Jones & Bartlett Learning, 1996.

[SL01] D. Seung and L Lee. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.

[SS97] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 1331–1334. IEEE, 1997.

[SSJ06] Roel Smits, Joan Sereno, and Allard Jongman. Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 32(3):733, 2006.

[Tan35] Arthur G Tansley. The use and abuse of vegetational concepts and terms. *Ecology*, 16(3):284–307, 1935.

[TLW10] Ran Tao, Yan-Lei Li, and Yue Wang. Short-time fractional fourier transform and its applications. *Transactions on Signal Processing*, 58(5):2568–2580, 2010.

[TMZ+06] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. *Cough*, 65(48):5, 2006.

[TN09] Andrey Temko and Climent Nadeu. Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, 30(14):1281–1288, 2009.

[TSGM10] Babak Taati, J. Snoek, D. Giesbrecht, and A. Mihailidis. Water flow detection in a handwashing task. In *Proceedings of the Canadian Conference on Computer and Robot Vision, CRV*, pages 175–182. IEEE, 2010.

[Tve77] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[VA12] Xavier Valero and Francesc Alías. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *Transactions on Multimedia*, 14(6):1684–1689, 2012.

[Van79] Nancy J. VanDerveer. *Ecological acoustics: Human perception of environmental sounds.* PhD thesis, Cornell University, 1979.

[Vap98] Vladimir N. Vapnik. *Statistical learning theory.* Wiley, 1998.

[VdD05] K. Van den Doel. Physically based models for liquid sounds. *ACM Transactions on Applied Perception, TAP*, 2(4):534–546, 2005.

[VIB+03] Michel Vacher, Dan Istrate, Laurent Besacier, Eric Castelli, and Jean-François Serignat. Smart audio sensor for telemedicine. In *Proceedings of Smart Object Conference*, pages 15–17, 2003.

[Vit67] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Transactions on Information Theory*, 13(2):260–269, 1967.

[VSN+11] Trang Thuy Vu, A. Sokan, H. Nakajo, K. Fujinami, J. Suutala, P. Siirtola, T. Alasalmi, A. Pitkanen, and J. Roning. Feature selection and activity recognition to detect water waste from water tap usage. In *Proceedings of the 17th International Conference on Embedded and Real-Time Computing Systems and Applications, RTCSA*, volume 2, pages 138–141. IEEE, 2011.

[Wan06] Avery Wang. The shazam music recognition service. *Communications of the ACM*, 49(8):44–48, 2006.

[WB+06] DeLiang Wang, Guy J Brown, et al. *Computational auditory scene analysis: Principles, algorithms, and applications*, volume 147. Wiley interscience, 2006.

[WCL+08] Chia-Chi Wang, Chin-Yen Chiang, Po-Yen Lin, Yi-Chieh Chou, I-Ting Kuo, Chih-Ning Huang, and Chia-Tai Chan. Development of a fall detecting system for the elderly residents. In *Proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering, ICBBE*, pages 1359–1362. IEEE, 2008.

[Wor08] Arthur Mason Worthington. *A study of splashes.* Longmans, Green, and Co., 1908.

[WV84] William H Warren and Robert R Verbrugge. Auditory perception of breaking and bouncing events: a case study in ecological acoustics. *Journal of Experimental Psychology: Human perception and performance*, 10(5):704, 1984.

[WW94] Dennis L. Wilson and James L. Wayman. Signal detector employing mean energy and variance of energy content comparison for noise detection, 1994. US Patent 5,323,337.

[YSD+12] Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatani, and Walter Kellermann. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. *Signal Processing Magazine*, 29(6):114–126, 2012.

[ZBT+09] Nadia Zouba, François Brémond, Monique Thonnat, Alain Anfosso, E Pascual, Patrick Mallea, Veronique Mailland, and Olivier Guerin. A computer system to monitor older adults at home: preliminary results. *Gerontechnology*, 8(3):129–139, 2009.

[ZCC+08] Zhongna Zhou, Xi Chen, Yu-Chia Chung, Zhihai He, T.X. Han, and J.M. Keller. Activity analysis, summarization, and visualization for indoor human activity monitoring. *Transactions on Circuits and Systems for Video Technology*, 18(11):1489–1498, 2008.

[ZDC06] Bin Zhang, Weibei Dou, and Liming Chen. Ball hit detection in table tennis games based on audio analysis. In *Proceedings of the 18th International Conference on Pattern Recognition, ICPR*, volume 3, pages 220–223. IEEE, 2006.

[ZHPHJ09] Xiaodan Zhuang, Jing Huang, Gerasimos Potamianos, and Mark Hasegawa-Johnson. Acoustic fall detection using gaussian mixture models and gmm supervectors. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 69–72. IEEE, 2009.

[ZJ09] C. Zheng and D.L. James. Harmonic fluids. In *Proceedings of Transactions on Graphics, TOG*, volume 28, page 37. ACM, 2009.

[ZWLH06] Tong Zhang, Jue Wang, Ping Liu, and Jing Hou. Fall detection by embedding an accelerometer in cellphone and using kfd algorithm. *International Journal of Computer Science and Network Security*, 6(10):277–284, 2006.