Ministère de l'enseignement Supérieur et de la recherche Scientifique
وزارة التعليم العالي والبحث العلمي

Badji Mokhtar Annaba University
Université Badji Mokhtar – Annaba
Faculté des Technologies
Département d'informatique

جامعة باجي مختار – عنابـــــــة
كلية التكنولوجيا
قســـــــم الإعلام الآلي

Thèse

Présentée pour obtenir le diplôme de Doctorat en Es-Sciences

Doctorat

Spécialité : Informatique

Par :
Abdoune Leila

Thème :

Séparation et classification des sources audio

N°	Nom et prénom	Grade	Etablissement	Qualité
01	FEZARI MOHAMED	Pr	Université Badji Mokhtar –Annaba	Directeur de thèse
02	SOUICI-MESLATI LABIBA	Pr	Université Badji Mokhtar –Annaba	Présidente
03	AZIZI NABIHA	Pr	Université Badji Mokhtar –Annaba	Examinateur
04	KOUAHLA MOHAMED NADJIB	Pr	Université 8 Mai 45 –Guelma	Examinateur
06	MAAZOUZI FAIZ	Pr	Université Mohamed Cherif Messaadia – SOUK AHRAS	Examinateur
05	CHEFROUR AIDA	MCA	Université Mohamed Cherif Messaadia– SOUK AHRAS	Examinateur

Année 2024-2025

Remerciements

Louange à *Allah*, pour m'avoir accordé la force, la patience et la volonté à l'aboutissement au terme de ce travail.

Je tiens à exprimer mes profonds remerciements à mon directeur de thèse Monsieur **Mohamed FEZARI**, Professeur à l'université Badji Mokhtar -Annaba - pour ses conseils, ses orientations, ses encouragements et son soutien permanent tant au niveau des connaissances qu'au niveau humain. Je le remercie pour ses efforts considérables et pour le temps qu'il a consacré à l'encadrement de ce travail de recherche.

Je tiens, également, à remercier les membres du jury qui m'ont fait l'honneur de bien vouloir évaluer mon travail, et plus précisément :

Mme la présidente **Souici-Meslati Labiba** Professeur à l'université Badji Mokhtar -Annaba pour l'honneur qu'elle m'a fait, en acceptant de présider ce jury, je la remercie pour ses encouragements, pour sa disponibilité et pour ses efforts.

Mme **Nabiha AZIZI**, Professeur à l'université Badji Mokhtar -Annaba, pour avoir accepté de faire partie de ce jury pour juger le présent document.

Monsieur **Mohamed Nadjib KOUAHLA**, Professeur à l'université 8 Mai 1945 -Guelma-, d'avoir accepté de faire partie de ce jury.

Mr **Faiz MAAZOUZI**, Professeur à l'université de Souk Ahras, pour avoir accepté de juger le présent document.

Mme **Aida CHEFROUR**, MCA à l'université de Souk Ahras, pour avoir accepté d'être membre de jury.

Je tiens à remercier tous mes collègues, enseignants, employés et responsables du département d'informatique.

Enfin, mes remerciements à tous les membres de ma famille qui m'ont tant soutenu, il s'agit en particulier de ma chère mère, et mon cher papa, mon cher époux qui a supporté mon absence durant toute ces années, mes sœurs et frères mes enfants : Nadine, Maher et Aya. Mes vifs remerciements aussi à ma belle-mère et mon beau père pour leurs encouragements et prières. Sans oublier mes amies et collègues qui m'ont toujours soutenu et encouragés : Asma, Fatma, fairouz, et Safia.

Dédicace

Je dédie cette réussite et le fruit de ces années d'effort à :

Mes parents, pour leur amour inconditionnel, leur soutien, leurs prières et leurs encouragements tout au long de ce parcours. Merci chère Maman et cher Papa pour votre patience et votre foi en moi.

Mon cher époux qui m'a toujours soutenu et supporté, merci pour ta patience et ta présence dans les moments les plus difficiles. Merci de m'avoir soutenu(e) et cru en moi même quand j'en doutais.

Mes enfants bien-aimés, Nadine, Maher et Aya, je dédie cette thèse avec l'espoir qu'elle vous inspire à suivre vos rêves et à croire en vos capacités.

Mes chers frères et sœurs : Zoubida, Fatiha, Saoudi, Nesma, Assia et Aimen, qui m'ont accompagné dans chaque étape de ma vie. Ce travail est autant le vôtre que le mien.

Ma belle-mère, mon beau père et ma belle-sœur Lynda, pour m'avoir fait sentir comme un membre à part entière de votre famille, pour votre soutien indéfectible, votre gentillesse et votre soutien, qui ont rendu ce voyage plus facile et plus enrichissant.

Ma chère amie d'enfance Dounia.

Safia, une personne extraordinaire qui est entrée dans ma vie au bon moment, qui, malgré notre rencontre récente, a su m'encourager et m'apporter une amitié précieuse. Merci pour ton incroyable soutien, tes paroles réconfortantes et tes encouragements inestimables.

Table des abréviations

Abréviation	Signification
ACP	l'Analyse En Composantes Principales
AReN	Audio Event Recognition Network
ASA	Auditory Scence Analysis
ASR	Automatic Sound Recognition
BDD	Base De Données
BER	Band Energy Ratio
BIC	Bayesian Information Criterion
CASA	Computational Auditory Scene Analysis
CFS	Correlation-Based Feature Selection
CNN	Convolutional Deep Neural Network
CRP	Cross Recurrence Plot
CWT	Continuous Wavelet Transform
DAG	Directed Acyclic Graph
DCT	Discrete Cosine Transform
DFB	Forward-Backward Divergence
DFT	Discrete Fourier Transform
DNN	Les Réseaux De Neurones Profonds
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
DWTC	Discrete Wavelet Transform Coeffcient
ECOC	Error Correcting Output Code
Er	Plage D'énergie
FDR	Fisher Descriminant Rtaio
FFT	Fast Fourier Transform
FS-P	Feature Selection—Perceptron
FWT	Fast (Discrete) Wavelet Transform
GTCC	Gammatone Cepstral Coefficients
GUI	L'interface Graphique
HCC	Homomorphic Cepstral Coefficients
HIS	Habitat Intelligent Pour La Santé
HMM	Hiden Markov Models
IoT	Internet Of Things
LSTM	Long Short-Term Memory
MFCC	Mel Frequency Cepstral Coefficients
MP	Matching Poursuit
mRMR	Minimum Redundancy Maximum Relevance
OAR	One-Against-The-Rest
PC	Personal Computer
PLP	Perceptual Linear Prediction
PR	Pitch Range

RAP	Reconnaissance Automatique De La Parole
RAS	Reconnaissance Automatique Des Sons
RASTA	Relative Spectral
REA	Reconnaissance Des Evénements Acoustiques
REA	Reconnaissance Des Evènements Audio
RFID	Radio Frequency Identification
RNA	Les Réseaux De Neurones Artificiels
RNN	Réseaux De Neurones Récurrents
Roll-off	Fréquence De Coupure
RSB	Le Rapport Signal Sur Bruit
RSE	Reconnaissance Des Sons Environnementaux
Sa	Asymétrie Spectrale
SC	Spectral Centroid
SED	Détection Des Evénements Sonores
SER	Reconnaissance D'événements Sonores
Sf	Planéité Spectrale
SIF	Spectrogram Image Feature
SNR	Signal To Noise Ratio
SPH	Single-Person Households
SRF	Spectral Roll-Off Point
STFT	Short-Time Fourier Transform,
SVM	Support Vector Machine
SVM-RFE	Recursive Feature Elimination For Support Vector Machines
Sw	Largeur De Bande Spectrale
TESPAR	Time Encoded Signal Processing And Recognition
VQ	Vector Quantization
WVD	Wigner-Ville Distribution.
ZCR	Zero Crossing Rate

Résumé

La reconnaissance des sons est devenue un domaine de recherche très actif ces dernières années, et elle soulève de nombreuses problématiques. Comme tout problème de reconnaissance, le choix des méthodes d'extraction de caractéristiques et de classification est l'un des problèmes les plus posés par ces systèmes car un bon choix de ces dernières influence positivement tout le système, et inversement. Le choix d'une méthode donnée dans un système de reconnaissance se fait soit selon des travaux antérieurs issus de l'état de l'art, ou bien par expérimentation en tenant compte des critères imposés par l'application visée ce qui conduit à enrichir l'état de l'art de nouveau.

Les domaines d'application de la reconnaissance des sons sont divers et multiples, et un point commun pour la plupart d'entre eux est la télésurveillance. En effet, la télésurveillance peut avoir lieu soit à l'intérieur, ou à l'extérieur, et c'est le premier cas qui nous intéresse et plus précisément la télésurveillance des personnes âgées ou handicapés vivant seules. La mise à disposition d'un système de télésurveillance audio pour cette catégorie est d'une grande importance et c'est l'objet de cette thèse. Par ailleurs, synthétiser des travaux antérieurs, étudier les sons qui peuvent avoir lieu dans un appartement, la recherche de méthodes appropriées pour la reconnaissance des sons et l'étude de la faisabilité des méthodes issues des domaines voisins, notamment la reconnaissance de la parole et de la musique, font aussi partie des objectifs principaux de ce travail.

En effet, Nous effectuons une étude comparative des travaux existants puis nous présentons l'architecture du système de reconnaissance des sons, mais avant de se faire, un corpus de sons de la vie courante est proposé. Les machines à vecteurs support (SVM) constituent notre premier centre d'intérêt vu leur puissance de séparation des différentes classes, combinés avec des paramètres acoustiques basés sur les MFCC (Mel Frequency Cepstral Coefficients). Le choix des SVM comme méthode de classification est le résultat d'une étude sur des travaux antérieurs où, vis-à-vis de notre application, le compromis entre la complexité des algorithmes et les performances du système est l'un des critères les plus notables. De plus, la plupart des applications pour la maison intelligente sont intégrées dans un produit matériel dont la puissance de calcul ne peut pas correspondre à celle d'un PC. Dans un premier temps, nous nous sommes intéressés à la construction d'une base de données pour les sons de la vie courante dans un habitat en partant du problème de l'absence d'une base de données standard. Cependant, dans un objectif de comparaison avec d'autres travaux une autre base de données a été utilisée. Les expérimentations réalisées dans cette thèse montrent l'efficacité des SVM et les MFCC pour la reconnaissance des sons de la vie courante qui sont de nature très diversifiés, et malgré que les tests effectués restent perfectibles vu la taille limitée de la base de données utilisée, ils sont encourageants et ouvrent la voie à plusieurs autres travaux et voire d'autres sujets de recherche.

Mots-clés : Reconnaissance des sons, télésurveillance, SVM, coefficients acoustiques, corpus de sons.

Abstract

The recognition of sounds has been a very active field of research in recent years, and it raises many issues. Like any recognition problem, the choice of methods of feature extraction and classification is one of the problems most posed by these systems, because a good choice of the latter positively influences the whole system and vice versa. The choice of a given method in a recognition system is made either according to previous work from the state of the art or by experimentation taking into account the criteria imposed by the intended application which leads to enriching the state of the art again.

The fields of application of sound recognition are diverse and multiple, and most of them have one thing in common is remote monitoring. The latter can take place either indoors or outdoors, and it is the first case that interests us, more specifically the remote monitoring of elderly or disabled people living alone. The provision of an audio monitoring system for this category is of great importance, and is the subject of this thesis. In addition, synthesizing previous work, studying the sounds that can occur in an apartment, finding appropriate methods for sound recognition and studying the feasibility of methods from related fields, notably speech and music recognition, are also among the main objectives of this work.

Indeed, we carry out a comparative study of existing works, then we present the architecture of the sound recognition system, but before doing so, a corpus of everyday life sounds is proposed. Support vector machines (SVM) are our first focus of interest due to their power of separation between different classes, combined with acoustic parameters based on MFCC (Mel Frequency Cepstral Coefficients). The choice of SVM as the classification method is a result of a study on previous work and depending on the type of our application where the trade-off between the complexity of the algorithms and the performance of the system is one of the most important criteria. As well as, taking into account that most smart home applications are built into a hardware product whose computing power cannot match that of a PC. Initially, we were interested in the construction of a database for everyday life sounds in a home, starting from the problem of a lack of a standard database. However, for the purpose of comparison with other works another database was used. The experiments conducted in this thesis demonstrate the effectiveness of SVM and MFCC in the recognition of everyday life sounds, which are highly diverse in nature. Although the tests carried out remain perfectible given the limited size of the database used, they are encouraging and pave the way for several other works and even other research topics.

Keywords: Sound recognition, remote monitoring, SVM, acoustic coefficients, sound corpus.

ملخّص

لقد أصبح التعرف على الأصوات مجالًا نشطًا جدًا للبحث في السنوات الأخيرة، وتوجد عدة اشكاليات في هذا المجال .مثل أي مشكلة في التعرف. يعد اختيار طرق استخراج الميزات وتصنيفها من أكثر المشكلات شيوعًا مع هذه الأنظمة، لأن الاختيار الجيد للأخير يؤثر بشكل إيجابي على النظام بأكمله والعكس صحيح .يتم اختيار طريقة معينة في نظام التعرف إما وفقًا للعمل السابق من حالة الفن أو عن طريق التجريب مع مراعاة المعايير التي يفرضها التطبيق المقصود والتي تؤدي إلى إثراء حالة الفن مرة أخرى.

مجالات تطبيق التعرف على الصوت متنوعة ومتعددة، ومعظمها يشترك في شيء واحد وهو المراقبة عن بعد .يمكن أن يحدث هذا الأخير إما في الداخل أو في الهواء الطلق وهذه هي الحالة الأولى التي تهمنا وبشكل أكثر تحديدًا المراقبة عن بعد للأشخاص المسنين أو المعاقين الذين يعيشون بمفردهم .إن توفير نظام مراقبة صوتي لهذه الفئة له أهمية كبيرة وهو موضوع هذه الرسالة .تجميع الأعمال السابقة، ودراسة الأصوات التي يمكن أن تحدث في شقة، والبحث عن الأساليب المناسبة للتعرف على الصوت، ودراسة جدوى الأساليب المستوحاة من المجالات المجاورة كالتعرف على الكلام والموسيقى أيضًا من بين الأهداف الرئيسية لهذا العمل.

في الواقع، نقوم بإجراء دراسة مقارنة للأعمال الموجودة ثم نقدم بنية نظام التعرف على الصوت، ولكن قبل القيام بذلك، يتم اقتراح مجموعة من الأصوات من الحياة اليومية .آلات ناقلات الدعم (SVM) هي محور اهتمامنا الأول نظرًا لقدرتها على الفصل بين الفئات المختلفة، جنبًا إلى جنب مع المعلمات الصوتية القائمة على) MFCC معامل تردد ميل التردد. ان اختيار SVM كطريقة تصنيف هو نتيجة دراسة أجريت على عمل سابق واعتمادًا على نوع تطبيقنا حيث تعد المفاضلة بين تعقيد الخوارزميات وأداء النظام أحد أهم المعايير .ناهيك عن أن معظم تطبيقات المنزل الذكي مدمجة في منتج أجهزة لا يمكن أن تتطابق قوتها الحاسوبية مع قوة الكمبيوتر. في البداية، كنا مهتمين ببناء قاعدة بيانات لأصوات الحياة اليومية في المنزل، بدءًا من فكرة عدم وجود قاعدة بيانات قياسية .ومع ذلك، تم استخدام قاعدة بيانات أخرى لغرض المقارنة مع الأعمال الأخرى .تظهر التجارب المنجزة في هذه الرسالة فعالية نماذج SVM و MFCC في التعرف على الأصوات البيئية التي تتميز بتنوع كبير، و على الرغم من أن الاختبارات التي تم إجراؤها ما تزال قابلة للتحسين نظرًا لحجم قاعدة البيانات المستخدمة، إلا أنها مشجعة وتمهد الطريق للعديد من الأعمال الأخرى وحتى الموضوعات البحثية الأخرى .

الكلمات المفتاحية:

التعرف على الصوت، المراقبة عن بعد، SVM، المعاملات الصوتية، مجموعة صوتية.

Table des matières

Table des illustrations

Liste des Tableaux

Introduction générale

1. Contexte de recherche

Avec l'évolution continue au fil des années des technologies d'information et de communication et les techniques d'intelligence artificielle il est presque devenu impossible de vivre sans l'outil informatique comme les ordinateurs, les Smartphones, les tablettes, et l'internet des objets. Ceci peut s'avérer dans différents domaines ; dans la vie quotidienne, le domaine médical, l'industrie, le commerce, etc. La maison intelligente constitue un bon exemple pour ces évolutions où on essaie de répondre aux besoins de l'habitant via l'assurance de divers services dont les plus importants la sécurité, la surveillance et l'assistance.

Le vieillissement de la population et l'augmentation du nombre des personnes âgées résidant toutes seules ainsi que le taux élevé des personnes atteintes de maladies chroniques telles que l'Alzheimer, les maladies cardiovasculaires et les grossesses à risque posent un problème important dans la société et même si nous pensons à maintenir toutes ces catégories dans les centres d'accueil ou dans les hôpitaux il sera impossible à cause de leur nombre élevé. Une des solutions offertes, ça fait une dizaine ou Vingtaine d'années, est de les suivre et surveiller dans leurs propres maisons via l'assurance de systèmes intelligents de télésurveillance qui peuvent informer les organisations spécialisées de leurs états et éventuellement de les informer en cas de détection d'une situation de détresse. Cela se fait par l'installation des capteurs dans la maison et plus particulièrement dans les différentes pièces de l'appartement (salon, cuisine, hall, toilette …) tels que les caméras, les microphones, les contacteurs de porte, l'infrarouge, les capteurs de présence, etc. A partir de ces différents capteurs, plusieurs types d'information peuvent être capturés : notamment l'activité effectuée, la localisation de la personne, et sa position qui peut être allongé, assis ou debout. Une fois ces informations sont renvoyées par le système installé et recueillies par l'équipe spécialisée, des décisions seront prises selon les informations acquises.

Plusieurs solutions ont été proposées dans la littérature, celles fondées sur la vidéo via l'utilisation des caméras dans les habitas, cependant cette solution a présenté deux inconvénients majeurs : le premier est qu'elle ne préserve pas davantage la vie privée des habitants, le deuxième est les problèmes liés aux caméras lorsqu'un obstacle se trouve dans le champ visuel ou lors d'un mauvais éclairage. D'autres solutions se basent sur la localisation des personnes via les cartes RFID (Radio Frequency Identification). Une autre solution adoptée est l'utilisation de canal audio dont l'objectif est la recherche de solutions moins couteuses, et avec une possibilité d'interagir avec l'environnement via les commandes vocales.

Notre travail est inspiré de [Dufaux, 2001] et [Cowling, 2004] qui tournent autour de la reconnaissance des sons pour une application de télésurveillance ou de reconnaissance des endroits. Notre environnement est l'habitat de la personne surveillée. En effet, afin de reconnaître les activités de l'habitant ainsi que la détection d'une éventuelle situation de détresse, plusieurs capteurs peuvent être installés dans l'habitat tels que les caméras, les microphones, les contacteurs de portes, l'infrarouge, l'accéléromètre, etc., chacun de ces capteurs peut apporter un type particulier d'information : l'emplacement de la personne, sa

position (allongé, debout), l'activité effectuée, etc. Afin de développer un tel système il est important de diviser le problème en sous problèmes chacun avec l'objectif à atteindre (reconnaissance des activités, reconnaissance des situations de détresse, ...). Lorsque l'ensemble des sous-systèmes sont développés nous pouvons arriver à un système complet qui peut répondre à tous les problèmes posés via la fusion des données résultantes de chaque sous système pour enfin arriver à une décision plus exacte.

Dans notre travail nous nous sommes intéressés par la détection des situations de détresse en utilisant le canal audio (microphones installés dans l'habitat). Notre système à concevoir est destiné pour la reconnaissance d'un nombre bien limité de classes des sons environnementaux, mais dans un entourage spécifique qui est l'habitat ou la maison. De ce fait, les sons environnementaux par lesquels nous nous sommes intéressés sont donc liés aux événements de la vie courante. La figure1 ci-dessous montre une vue globale du système de reconnaissance à réaliser et positionne notre travail par rapport au système globale.

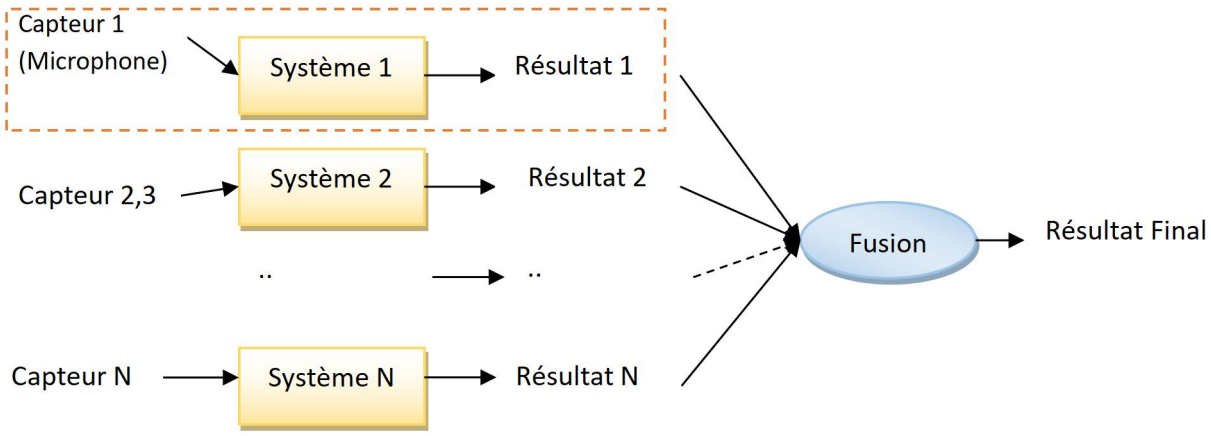

Figure 1. Vue globale du système Final et le sous-système de reconnaissance de sons

En effet, le contexte de ce travail s'articule autour de deux axes principaux : les maisons intelligentes et l'assistance aux personnes âgées.

Le terme *maison intelligente* ou « *smart home* » a été officiellement utilisé en 1984. La *maison intelligente* d'après [Harper, 2003] est un habitat équipé de technologies d'informations et de communication conçus pour collaborer afin d'anticiper et de répondre aux besoins des occupants, en travaillant pour promouvoir leur confort, sécurité et divertissement tout en préservant leur interaction naturelle avec l'environnement. La maison intelligente est une branche importante de la *domotique* [Vacher, 2011].

L'objectif principal des maisons intelligentes est de donner aux occupants un contrôle total sur la maison de n'importe où dans la maison ou d'un endroit éloigné [DiCarlo et Cove, 2009]. Mais, en réalité les objectifs des maisons intelligentes vont plus loin que ça. Bonhomme dans sa thèse [Bonhomme, 2008] a divisé les objectifs des habitats intelligents en deux grandes catégories :

- *La sécurité des biens et des personnes* : via les services de surveillance, services de soins, téléassistance, assistance thérapeutique, assistance médicale, etc., ces services permettent de veiller sur les personnes ayant des handicaps moteurs, visuels, auditifs ou cognitifs ainsi que sur les personnes âgées, dans le cadre du maintien à domicile.

- *La gestion du confort :* qui consiste à optimiser le confort, le bien-être et la qualité de vie de ses habitants, en assurant à la fois le confort d'usage par le multimédia et le confort sensoriel par régulation de l'ambiance, gestion énergétique, etc.

Plusieurs études montrent que le nombre de capteurs et de dispositifs de la maison intelligente du future va augmenter et ceci est de l'ordre de 15 à 30 dispositifs et capteurs connectés via un réseau aux fournisseurs de services et à internet [Chan et al., 2008]. Par conséquent, Les objectifs des maisons intelligentes vont aussi augmenter. Les principaux services auxquels la maison intelligente va s'occuper sont : services d'utilité (électricité, gaz, eau…), services santé (home health), services de sécurité et services de divertissement. Ainsi, les maisons intelligentes ambitieuses peuvent être examinées du point de vue du confort, des loisirs et de la sécurité.

Parmi les points et les changements qui ont favorisé l'évolution des maisons intelligentes d'après [Chan et al., 2008] et [Bonhomme, 2008] :

- La disponibilité des ordinateurs personnels (PC) avec un accès à Internet.
- Evolution de la technologie des capteurs ; taille plus réduite, autonomes et sans fil.
- Les micro-objets passifs ou actifs 'tags' sont de plus en plus utilisés. Ce sont des puces embarquées, des « étiquettes » avec lesquelles il est possible de communiquer.
- Les téléphones portables qui facilitent la communication et les interactions entre les habitants (personnes surveillées) et le monde extérieur (les différents intervenants).

Une application importante de l'habitat intelligent est l'Habitat Intelligent pour la Santé (HIS) ou 'Health Smart Home' en anglais [Vacher et al., 2010a]. Dans le domaine médical, les grands défis auxquels l'HIS contribue sont : l'hospitalisation à domicile, la télémédecine et la téléassistance sociomédicale. Les populations visées en priorité sont les personnes âgées ou les personnes isolées, les patients chroniques (handicapés physiques, cardiopathes) et les personnes sous surveillance médicale temporaire (grossesses à risque, patients en postopératoire, etc.).

La téléassistance des personnes âgées remonte aux années 80, où les personnes peuvent bénéficier d'un service de maintien et d'assistance à domicile à travers un dispositif électronique qui permet de mettre la personne en contact avec sa famille ou un service spécialisé en cas de danger par une simple pression sur un bouton.

Aujourd'hui, L'assistance des personnes âgées est considérée comme un des objectifs et défis majeurs de la maison intelligente, pour cette raison plusieurs recherches se concentrent sur la conception et la mise en œuvre de maisons intelligentes pour une population bien spécifique, celle des personnes âgées vivant seules, mais aussi les handicapés, et les personnes atteintes de

maladies spécifiques tels que l'Alzheimer et les maladies chroniques. Plusieurs projets ont été lancés pour l'assistance des personnes âgées nous citons : SweetHome [1] qui vise à améliorer l'autonomie, le confort et la sécurité à la maison en utilisant une commande vocale intégrée à un système domotique standard pour interagir avec l'environnement, PROSAFE [2], [Chan et al., 2003] pour la surveillance continue des personnes âgées ou handicapées à domicile, RESIDE-HIS [Istrate et al., 2006] qui a comme objectif la détection d'une situation de détresse de la personne, Homecare [3] qui vise à expérimenter et à qualifier un système complet de télésurveillance pour les personnes âgées atteintes de la maladie d'Alzheimer.

2. Motivations

La reconnaissance des sons est un domaine moins étudié en comparaison avec d'autres domaines tels que la reconnaissance de la parole et la musique. Si l'objectif de l'intelligence artificielle est de doter la machine par les fonctions sensorielles de l'être humain alors, en comparaison avec la recherche dans le domaine de la vision, la fonction d'audition est moins explorée. Par conséquent, il est très important de s'investir dans le domaine de la reconnaissance des sons environnementaux. Outre l'intérêt purement scientifique, plusieurs domaines peuvent bénéficier de la reconnaissance de sons tels que la robotique, les technologies d'aide auditives, les systèmes de sécurité et aussi les systèmes de télésurveillance et d'assistance.

Le vieillissement de la population dans le monde et aussi le changement du mode de vie dans presque tous les pays où après un certain temps les parents vivaient tout seuls et si l'un du couple meurt l'autre passe le reste de sa vie tout seul chose qui nécessite de trouver des solutions pour prendre soins de cette catégorie que ce soit au niveau sociologique en réduisant par exemple leur solitude ou au niveau pratique par assurance d'outils pour leur faciliter la vie.

Enfin, L'émergence du concept *'maisons intelligentes pour la santé'* (HIS) qui sont conçues pour soutenir la vie quotidienne afin de compenser certaines incapacités chez l'habitant tels que l'oubli, les difficultés de l'ouïe, et aussi pour la détection de situations de détresse constitue une des motivations principales pour ce travail.

3. Objectifs et problématiques de la thèse

L'objectif de cette thèse est l'étude et la validation d'un système de classification des sons de la vie courante pour la détection d'une situation de détresse. Les sons étudiés sont des sons qui peuvent être acquis dans un habitat, parole ou autre son, et ils peuvent indiquer soit un état normal ou de détresse de l'habitant. La catégorie de la population visée est les personnes âgées ou handicapées. Ce système fait partie des systèmes de *télésurveillance* via le canal audio et aussi fait partie des projets et objectifs de *la maison intelligente.* La plupart des systèmes de télésurveillance existants utilisent les capteurs vidéos mais nos recherches sont orientées vers l'utilisation des capteurs sonores qui sont moins coûteux, n'affectent pas la vie privée des habitants, et aussi peuvent être utilisés en complément avec des solutions basées sur les caméras lorsque ces dernières ne peuvent pas capter les informations désirées à cause d'un

obstacle dans le champ visuel ou d'un mauvais éclairage. Il est aussi important de comparer l'information visuelle et audio du point de vu stockage et calcul, les signaux audio utilisent moins d'espace mémoire et puissance et temps de calcul par rapport à l'information visuelle [Chu et al., 2009].

Le domaine de la reconnaissance des sons est un domaine assez complexe ceci est dû à plusieurs facteurs, nous citons :

- Le nombre important des sons qui peuvent se produire dans un endroit donné. Plus le nombre de classes de son augmente plus le système devient de plus en plus compliqué. Par conséquent, La plupart des travaux se concentrent sur un type particulier de sons.
- La diversité des sons et leurs structures différentes : stationnaires, non stationnaires, quasi stationnaires et impulsifs et l'aspect non stationnaires de la plupart des sons environnementaux qui rend leur modélisation une tâche difficile.
- Le bruit environnemental reste un des problèmes majeurs de la reconnaissance de sons.
- Le domaine est récent et le nombre de recherches est petit par rapport à la reconnaissance de la parole/locuteur et la musique.

Les problématiques de cette thèse sont diverses, nous citons :

- *Absence d'un corpus de son de la vie courante* : la première phase de notre travail consiste à collecter les données et plus particulièrement la définition d'un corpus de son de la vie courante en définissant les classes de sons qui peuvent être générés dans l'environnement restreint maison.
- *Le volume important des informations* : plusieurs sons peuvent être acquis en même temps sur les différents canaux audio.
- *La présence du bruit environnemental* : l'environnement maison est un milieu qui peut contenir une variété de sons comme il peut comporter le bruit aussi. Par conséquent, il faut tenir compte de l'aspect bruit environnemental dans ce travail.
- *Difficulté de validation du système proposé* en se basant sur les résultats obtenus à cause de l'absence d'une base de données commune : la comparaison avec d'autres systèmes ne peut pas se faire de manière exacte car la majorité ou presque tous les systèmes de reconnaissance des sons évaluent leur systèmes sur leur propres données d'où la difficulté de définir les paramètres acoustiques les plus efficaces ou appropriés pour la reconnaissance des sons ainsi que les techniques de classification les plus adéquates.
- *Compromis entre la complexité des algorithmes et les performances du système* : le caractère de l'application visée qui est la télésurveillance des personnes âgées ou handicapées et son aspect temps réel nécessitent des solutions avec un temps de calcul réduit mais en garantissant de bonnes performances.

L'objectif de ce travail se concentre autour des points et enjeux suivants :

- Présenter et étudier le domaine de la reconnaissance des sons tant au niveau paramètres acoustiques et méthodes d'extraction de caractéristiques qu'au niveau méthodes de classification.
- En se basant sur une synthèse des travaux sur la reconnaissance des sons et les systèmes aussi de détection de situations de détresse via le canal audio, trouver les méthodes d'extraction de caractéristiques et de classifications les plus appropriées pour la reconnaissance des sons.
- Proposition d'un corpus de sons pour la vie courante.
- Evaluer le système final et discuter les résultats.
- Mise en œuvre via l'utilisation d'un microcontrôleur.

Dans ce travail, nos expérimentations ont commencé avec l'utilisation des MFCC (Mel Frequency Cepstral Coefficients) comme paramètres acoustiques et un classifieur basé SVM (Support Vector machines). En effet, nos choix sont justifiés par nos études et synthèses et un des objectifs de cette étude qui est laissé en perspectives est de tester d'autres paramètres acoustiques autres que les MFCC et leurs dérivée première et deuxième, et même de les combiner afin d'améliorer les taux de reconnaissance.

Afin de mettre en relief nos contributions qui constituent une bonne partie des objectifs visées par cette thèse, citées précédemment, nous les énumérons comme suit :

- Étude approfondie du domaine de la reconnaissance de son en commençant de la définition des sons passant par sa numérisation puis en se focalisant sur les taxonomies de sons existante et enfin description des méthodes d'extraction de caractéristiques et de classification qui sont liées et utilisées dans la reconnaissance des sons.
- État de l'art et synthèse des travaux sur la reconnaissance des sons en vue de rechercher les méthodes les plus appropriées pour le domaine de la reconnaissance des sons environnementaux non parole. Cette étude a été suivie aussi par une synthèse des travaux liés à la détection des situations détresse dans le but toujours de voir quelles méthodes sont utilisées, quelles classes sont visées et quels sont les taux de reconnaissance obtenus.
- Vu les conclusions tirées du premier et deuxième point de notre contribution qui est le choix de nos méthodes d'extraction de caractéristiques et de classification, nous avons fait un aperçu sur les travaux liés à la reconnaissance des sons, mais qui se basent sur les méthodes d'apprentissage profond afin de monter leur efficacité et en les comparant avec des méthodes d'apprentissage classiques tel que les SVM.
- Nous avons consacré un chapitre à part pour la construction d'un corpus de sons de la vie courante et initialement la construction de la base de données qui a fait l'objet d'un premier travail publié [Abdoune et Fezari, 2016]. L'objectif de cette BDD (Base de Données) est pour le test et la validation du système final, mais pour des raisons bien expliquées dans ce manuscrit nous avons opté pour une autre base de données dans nos expérimentations.

- Dans un autre point de notre travail, nous avons d'abord présenté différentes architectures pour la reconnaissance et la classification des sons qui sont considérés un cocktail de parole, musique et sons environnementaux. Nous avons ensuite présenté notre architecture inspirée des architectures étudiées.
- Enfin, une évaluation de notre classifieur basé SVM sur un nombre de classes bien choisi, puis interprétation et discussion des résultats.

4. Organisation de la thèse

Au vu de nos objectifs et problématiques posées, nous avons organisé ce manuscrit de thèse de la manière suivante en plus de cette introduction générale et d'une conclusion.

Chapitre 1 : décrit la théorie de base de la reconnaissance de sons. Nous nous focalisons sur les concepts de base de la reconnaissance des sons, les méthodes d'extraction de caractéristiques et les méthodes de classification.

Chapitre2 : dans ce chapitre, nous décrivons les travaux et les méthodes utilisées pour la reconnaissance des sons de l'environnement. Ensuite, nous abordons la reconnaissance des sons dans un contexte domotique et en particulier pour la détection des situations de détresse dans un habitat. L'idée de base de ce chapitre est de donner un état de l'art sur le domaine de reconnaissance des sons et les solutions proposées.

Chapitre3 : constitue notre première contribution, il aborde et décrit le corpus de sons de la vie courante, les classes de sons, les conditions et paramètres d'enregistrement, etc. Pour cette raison des travaux antérieurs sur la création des bases de données ainsi que les travaux utilisant des bases de données existantes sont présentés au début de ce chapitre. Une expérimentation est décrite dans la dernière partie de ce chapitre pour une première validation de notre base et corpus de sons de la vie courante.

Chapitre 4 : présente notre deuxième contribution où nous décrivons l'architecture générale d'un système de reconnaissance des sons de la vie courante ainsi que l'architecture détaillée du système de reconnaissance des sons de l'environnement maison. Ce chapitre décrit notre expérimentation basée sur l'SVM et les paramètres acoustiques MFCC avec une interprétation et discussion des résultats.

Conclusion : Cette partie décrit et présente une courte synthèse du travail réalisé dans cette thèse, en rapporte les conclusions, et identifie les voies de recherche qui nous semblent les plus prometteuses**.**

CHAPITRE 1
Reconnaissance de son

Avant de s'attaquer à notre problématique et avant même de montrer les solutions existantes et les approches et méthodes utilisées, il est impératif de présenter d'abord la théorie de base de la reconnaissance de sons, et c'est l'objet de ce premier chapitre. Nous nous focalisons sur la théorie de la reconnaissance des sons, les méthodes d'extraction de caractéristiques et les méthodes de classification.

1.1. INTRODUCTION

Dans la vie quotidienne et dans à n'importe où, nous rencontrons une variété de sons comme les chants d'oiseaux, les cris, les sons de vaisselles, sonnerie téléphone et les sons de pas. Cet ensemble de sons varie d'un endroit à un autre et varie même dans le même endroit selon le scénario vécu où les activités effectuées. Pouvoir détecter et reconnaitre automatiquement ces événements sonores s'appelle **reconnaissance automatique des sons** (RAS). Dans ce travail, on s'intéresse à la **reconnaissance des sons environnementaux** (RSE). D'autres termes aussi peuvent être utilisés et indiquant le même sens que la reconnaissance des sons tel que **reconnaissance des événements acoustiques** (REA), **reconnaissance des évènements audio** (REA) ou **reconnaissance des sons de la vie courante.** L'ensemble de ces termes fait partie ou indique le domaine de la **reconnaissance du son** et ils sont utilisés dans ce manuscrit.

La majorité des recherches se sont concentrées sur la reconnaissance de la parole et du locuteur, cependant les travaux sur la reconnaissance des sons de l'environnement non parole sont très peu, mais qui sont devenu plus actifs ces dernières années.

En réalité, la reconnaissance des sons est un domaine qui combine entre les méthodes de l'intelligence artificielle et les méthodes de traitement de signal. Pour cette raison, il est important pour pouvoir comprendre le principe de la reconnaissance des sons de traiter dans ce chapitre les deux parties de ce domaine. Ce chapitre est organisé ainsi : la première partie s'agit de la théorie de base de la reconnaissance de son à savoir : le son, sa numérisation, ses catégories et taxonomies, puis le processus de reconnaissance. Une deuxième partie est consacrée pour les méthodes d'extraction des caractéristiques acoustiques et les méthodes de classification des sons. Pour les méthodes d'extraction des caractéristiques nous nous focalisons sur les MFCC et pour les méthodes de classifications nous mettons l'accent sur les SVM vu que ces deux dernières méthodes seront utilisées dans notre système.

1.2. RECONNAISSANCE DE SON : ANALYSE DE L'EXISTANT

Comprendre le principe de la reconnaissance des sons et pouvoir localiser ses méthodes est l'un des objectifs majeurs de cette thèse. En effet, le terme son peut signifier et contenir plusieurs autres types, par exemple la parole est un son, la musique l'est aussi, les sons environnementaux comme la pluie et le tonnerre sont un autre type de son. De ce fait, il est important de mettre en avant les différences clés et de préciser quel type de son est ciblé par cette thèse.

Dans cette section nous examinerons la théorie de base de la reconnaissance de son qui est nécessaire pour comprendre le domaine de recherche. Afin de comprendre comment se déroule la reconnaissance de sons, nous voyons important de commencer par définir l'entité principale du système qui est le son et comment ce dernier peut être numérisé, puis nous décrivons les catégories de sons ensuite nous présentons les taxonomies des sons en essayant d'en adopter une selon notre point de vue. Finalement, nous décrivons de manière globale le processus de reconnaissance de sons et en détaillant chaque module à part.

1.2.1. Qu'est-ce qu'un son ?

Le son est généré lorsqu'un objet provoque une perturbation de la densité du milieu qu'il abrite (où il réside) [Tipler, 1991]. Cette perturbation se propage à travers le milieu et lorsqu'elle atteint l'oreille de l'être humain, elle est convertie en des signaux électriques que le cerveau interprète comme son. La perturbation produite est sous forme d'une onde (figure 1.1), avec une amplitude qui signifie le nombre des mouvements des molécules, et une fréquence signifiant la durée ou l'intervalle de temps avant laquelle l'onde (waveform) se répète.

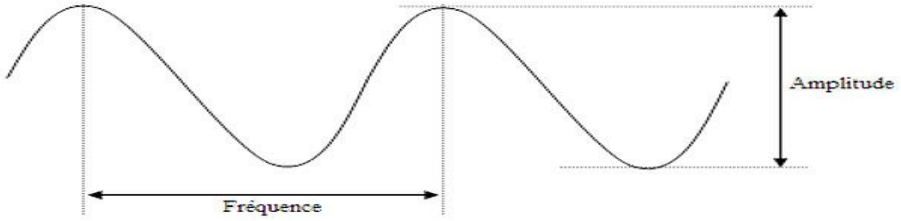

Figure 1. 1. Une onde

1.2.2. Numérisation d'un signal

1.2.2.1. Echantillonnage

Un signal analogique peut être vu comme une onde ou une suite d'ondes continues. Lorsque ce signal est stocké dans l'ordinateur il est transformé en un signal numérique en prenant les valeurs de l'onde dans des intervalles réguliers. Le signal continu est alors transformé en une suite de valeurs discrètes. Nous appelons cette étape *l'échantillonnage* (figure 1.2)

Les deux termes fréquence d'échantillonnage ou taux d'échantillonnage f mesuré en Hz or kHz, désignent le nombre d'échantillons par unité de temps. Si l'unité de temps est la seconde, la fréquence d'échantillonnage s'exprime en hertz et représente le nombre d'échantillons utilisés par seconde.

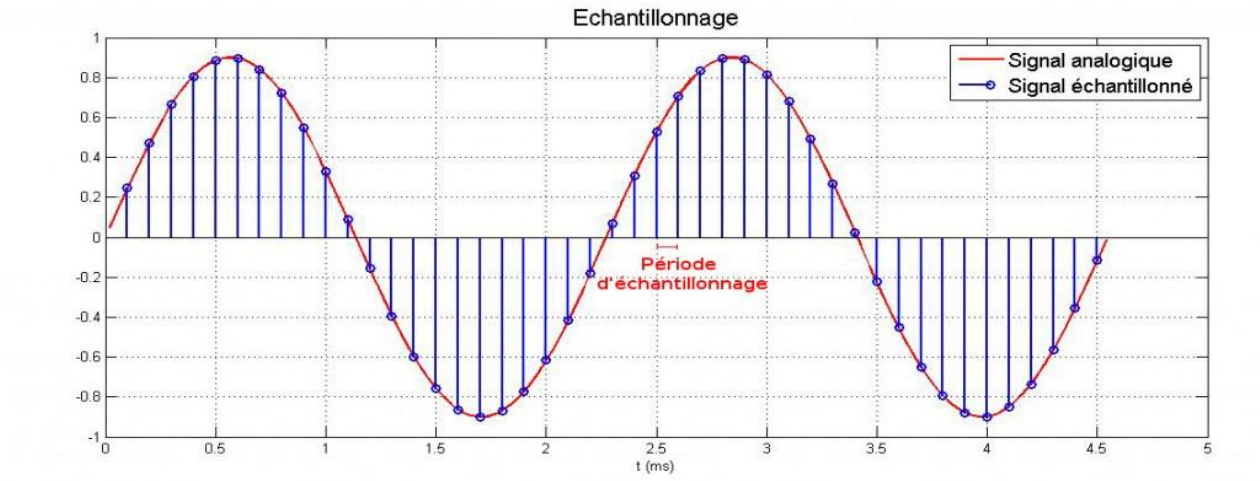

Figure 1. 2. Echantillonnage d'un signal

1.2.2.2. Quantification

La quantification est une étape qui suit l'échantillonnage (figure 1.3 et figure 1.4), elle consiste à stocker l'amplitude du signal à chaque point d'échantillonnage [Kefauver, 1999]. La quantification consiste à associer une valeur numérique pour chaque échantillon selon son amplitude. Ces valeurs numériques sont attribuées selon une échelle de bits [Sueur, 2018]. Une quantification de 8 bits assignera des valeurs d'amplitude sur une échelle de $2^8 = 256$ états autour de 0 (zéro). La quantification de 16 bits est la plus utilisée dans la plupart des systèmes d'enregistrement.

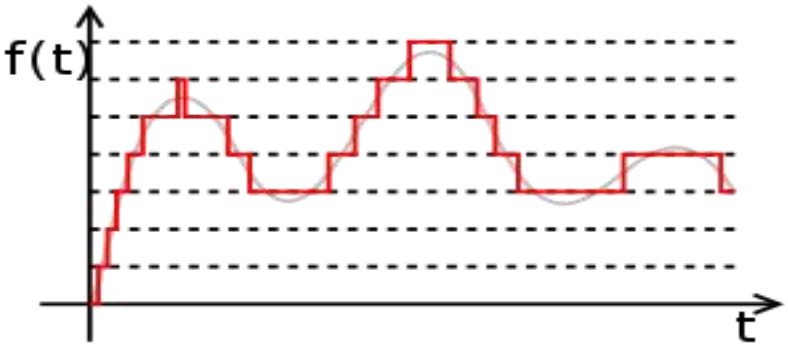

Figure 1. 3. Quantification d'un signal

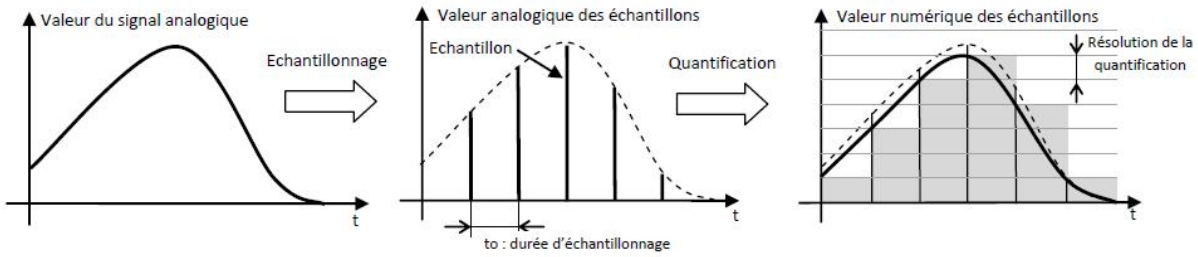

Figure 1. 4. Echantillonnage et quantification

1.2.2.3. Fenêtrage et transformations

Après l'étape d'échantillonnage et de quantification, différents traitements peuvent être subis par le signal numérisé, le plus souvent sont les transformées et le fenêtrage.

Les transformées sont les plus importants traitements qui peuvent être appliqués sur un signal et en particulier la transformée de Fourier [Shie et Chen, 1999], [Cowling, 2004]. Cette méthode consiste à représenter les échantillons du signal par des séries de Fourier pour simplifier les calculs, en décomposant une onde complexe en des composantes sinusoïdales. Il existe d'autres transformations telles que la transformée de Laplace, la FFT (Fast Fourier Transform), la DFT (Discrete Fourier Transform), l'expansion de Gabor et la STFT (Short-Time Fourier Transform), mais elles ont tous le même principe de la transformée de Fourier avec quelques différences. La FFT par exemple, est une implémentation optimisée de la DFT et consomme moins de temps de

calcul, elle transforme la représentation du signal du domaine temporel vers le domaine fréquentiel pour une analyse ultérieure des fréquences du signal.

En effet, dans le domaine de la reconnaissance de son une alternative pour la transformée de Fourier est la DCT (Discrete Cosine Transform) qui supprime quelques composants haute-fréquence erronés qui sont introduits dans le spectre (spectrum) avec une FFT tel qu'expliqué dans [Cowling, 2004]. De plus, la STFT est considérée comme une bonne méthode pour l'analyse des signaux comme les signaux impulsifs, qui ont un spectre de fréquence qui varie dans le temps [Dufaux, 2001].

En plus des transformées existantes, *le fenêtrage* peut être aussi appliqué sur le signal afin d'améliorer la reconnaissance des différentes formes existantes dans le signal. Plusieurs techniques de reconnaissance de la parole/locuteur utilisent les fenêtres en chevauchement d'un signal pour améliorer la reconnaissance [Cowling, 2004].

Le fenêtrage est la multiplication d'un signal par une fenêtre de longueur finie dont l'amplitude varie doucement et progressivement vers zéro sur les bords. C'est une technique qui est utilisée pour minimiser les effets d'application de la FFT sur un nombre non entier de cycles [5]. Cette technique sert à réduire l'amplitude des discontinuités sur les bornes de chaque séquence finie acquise par le numériseur.

Il existe plusieurs types de fenêtres (figure 1.5) tels que la fenêtre de Hamming, fenêtre de Hann, fenêtre de Blackman. La plus simple et standard est la fenêtre rectangulaire, et la fenêtre la plus utilisée en reconnaissance de la parole est la fenêtre de Hamming.

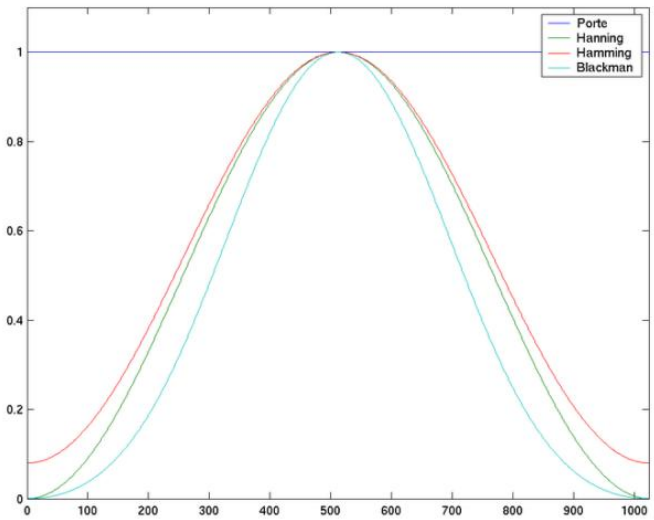

Figure 1. 5. Exemple de fenêtres : fenêtre rectangulaire, fenêtre de Hanning, fenêtre de Hamming et fenêtre de Blackman

1.2.3. Les catégories de sons

Dans cette section nous décrivons les différents types de sons et leur classification en essayant de montrer les principales différences. En effet, un son peut être de type parole, musique ou un son environnemental, pour chacun de ces trois types un son peut être impulsionnel, stationnaire ou non stationnaire. Avant de procéder au traitement des données en utilisant les techniques existantes il faut d'abord comprendre la nature des données et par conséquent donner un aperçu pour les différentes catégories de sons qu'on peut avoir dans la vie quotidienne et même descendre à un niveau plus bas pour décrire les propriétés du signal de ces catégories de son.

1.2.3.1. La nature des signaux audio de la musique, parole et sons environnementaux

Il existe trois catégories des sons *: la parole, la musique et les sons de l'environnement*. En effet, chaque catégorie de sons possède des caractéristiques distinctes et aussi des catégories différentes peuvent avoir des caractéristiques communes et ceci dépends du type de son étudié de chacune de ces catégories. Chachada et Kuo dans [Chachada et Kuo, 2013] définissent les sons environnementaux par les sons quotidiens (naturels ou artificiels) autres que la musique et la parole. Goldhor dans [Goldhor, 1993] définit les sons environnementaux comme les sons générés par des sources acoustiques dans les environnements domestiques, professionnels et extérieurs. Par conséquent, de notre point de vue, l'ensemble de ces trois catégories de son nous frome ce que nous appelons *'sons de la vie courante'* qui est un mélange de la parole et n'importe quel autre type de son (figure 1.6).

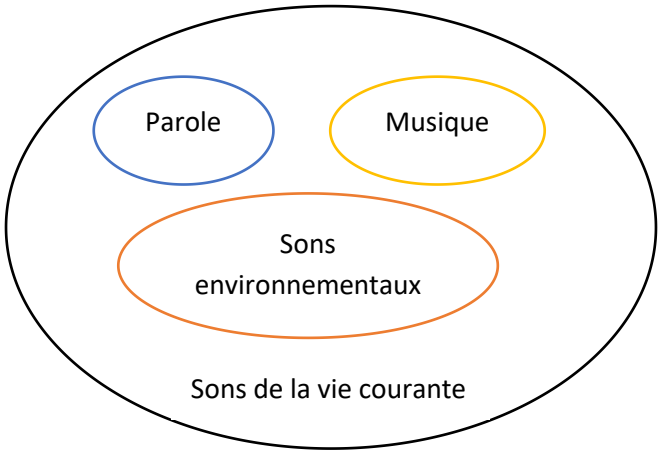

Figure 1. 6. Sons de la vie courante

Dans cette partie, nous présentons un aperçu sur ces trois catégories de sons en mettant en relief les différences clés.

1) La parole

Les signaux de parole sont des sons produits par l'appareil vocal impliquant un contenu linguistique. La parole présente des traits distinctifs qui la différencient des autres types de sons

partant de sa distribution spectrale caractéristique à sa structure phonétique [Alías et al., 2016]. Plusieurs travaux sur la parole existent dans la littérature tels que l'identification et la reconnaissance du locuteur [Kinnunen et Li, 2011], [Chung et al., 2018], la reconnaissance de la parole [Pieraccini, 2012], [Xiong et al., 2018], la détection de la parole [Bach et al., 2011] ainsi que la détection de discours de haine [Watanabe et al., 2018] qui présente un nouveau domaine de recherche pour la reconnaissance de la parole ces dernières années. La parole utilise la structure phonétique, i.e., chaque morceau de parole peut être vu comme un ensemble de *phonèmes* qui peuvent être modélisés par des HMM (Hiden Markov Models), ce qui n'est pas le cas pour les signaux environnementaux.

2) La musique

La musique est aussi un son *structuré* comme c'est le cas pour la parole et présente aussi un ensemble de traits qui la caractérisent. Elle présente des modèles ou formes *stationnaires* telles que *la mélodie* et *le rythme* contrairement aux sons environnementaux. Différents types de systèmes ont été développés qui traitent les sons de type musique tels que la reconnaissance des instruments [Han et al., 2017], [Gururani et al., 2018], classification de genre [Kour et Mehan, 2015], [Zhang et al., 2016], identification des chansons et artistes [Ratanpara et Patel, 2015], annotation musicale et recommandation [Lyon, 2011], et classification de l'humeur [Lu et al., 2006], [Ren et al., 2015], [Sarno et al., 2018].

3) Les sons environnementaux

Contrairement à la parole et la musique, les sons de l'environnement ne présentent pas des traits spécifiques. Cependant, cette catégorie de son doit être reconnue par les systèmes d'audition comme pour la parole et la musique, soit comme des sons individuels [Vacher et al., 2010a], [Vacher et al., 2010b]., soit comme des scènes acoustiques [Chu et al., 2006], [Chu et al., 2009], [Uzkent et al., 2012].

En général, les sons environnementaux se caractérisent par le fait qu'ils ne sont pas stationnaires, en plus ils ne possèdent pas la structure des phonèmes ce qui rends la reconnaissance une tâche difficile. De plus, même si on peut décomposer le son en sous unités qui ressemblent aux phonèmes il est difficile de modéliser leur variation dans le temps par les HMM et leurs occurrences temporelles sont aléatoires comme les sons de tonnerre [Chachada et Kuo, 2013].

4) Sons de la vie courante

Nous désignons par les sons de la vie courante tout type de son qui peut être généré dans l'environnement, soit à l'intérieur de la maison ou au bureau ou à l'extérieur. Les sons de la vie courante peuvent être de la parole, la musique, ou tout autre son (télévision, claquement, ouverture ou fermeture de portes, clacksons de voitures, pluie, etc.).

La nature de l'application précise l'environnement d'intérêt, dans notre cas l'environnement d'intérêt c'est l'habitat ou la maison. Une description plus détaillée pour les sons de la vie courante est présentée dans le chapitre 3. Donc, un système de reconnaissance des sons de la vie courante doit tenir compte des caractéristiques des différents types de sons qu'il englobe ce qui nous mène à penser à trouver des solutions.

1.2.3.2. Types de sons vis-à-vis des Caractéristiques du signal audio

Après présentation des différentes catégories de sons existantes musique, parole et sons de l'environnement, une nouvelle classification de ces trois catégories peut se faire selon la nature du signal extrait du son mais pas sur la catégorie de son. Un signal sonore peut être impulsif, stationnaire et non stationnaire. En plus, une autre catégorie intermédiaire entre le stationnaire et non-stationnaire peut être ajoutée c'est le quasi-stationnaire. Dans ce qui suit, nous présentons ces différents concepts.

1) Les sons impulsifs

Le son impulsif est un son qui dure pendant une courte période et comprend des fréquences couvrant une grande partie du spectre acoustique, tel qu'un coup de marteau ou un claquement de main [4]. Des exemples sur les sons impulsifs sont : coups de feu, claquement de porte et cris.

Il existe des travaux de reconnaissance des sons impulsifs telle que la thèse de Dufaux [Dufaux, 2001]. D'après lui, un bon système de reconnaissance des sons impulsifs doit tenir compte des propriétés non stationnaires du signal et aussi en considérant les dynamiques temporelles. Un autre travail présenté par [Arslan, 2017] consiste en la détection et la reconnaissance des sons impulsifs et en particulier les coups de feu, il propose une nouvelle méthode pour la détection des sons impulsifs, en présentant une nouvelle formule d'énergie prenant en compte la moyenne et la variance de la séquence du signal.

Les sons impulsifs présentent donc une partie des sons environnementaux qui sont caractérisés par leur intensité et leur courte durée.

2) Qu'est-ce qu'un signal stationnaire, quasi-stationnaire et non-stationnaire ?

On parle d'un son stationnaire, quasi stationnaire ou non stationnaire selon sa vitesse de variation dans le temps, c'est le cas des sons non paroles qui peuvent être l'un de ces trois types. En effet, les sons stationnaires ne contiennent pas des changements rapides dans leur spectre à travers le temps. Les sons quasi-stationnaires ont principalement un spectre constant dans le temps. Les sons non stationnaires contiennent de larges et rapides changements dans le spectre au fil du temps. Pour les sons stationnaires, l'utilisation des méthodes ou techniques de traitement de signal simples, tels que l'FFT peuvent être suffisantes pour la reconnaissance des sons stationnaires. En revanche, il est difficile de reconnaitre les sons quasi-stationnaires et non-stationnaires vu les changements rapides au niveau de leurs caractéristiques.

- **Les sons stationnaires**

Un son stationnaire est un signal avec des fréquences contenues également tout au long de l'ensemble du signal [Cowling, 2004]. Nous citons à titre d'exemples : le bruit blanc, bourdonnement d'un ventilateur, moteur ou générateur électrique fonctionnant à un régime constant (voir figure 1.7).

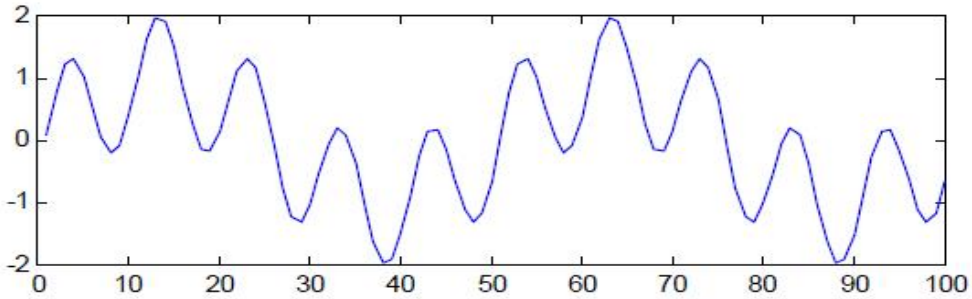

Figure 1. 7. Un signal avec des fréquences constantes (stationnaire)

- **Les sons non stationnaires**

Les sons non-stationnaires (figure 1.8) ne contiennent pas les mêmes fréquences dans toutes les parties du signal [Cowling, 2004]. Nous donnons comme exemple, les sonneries de téléphone, les cloches, etc.

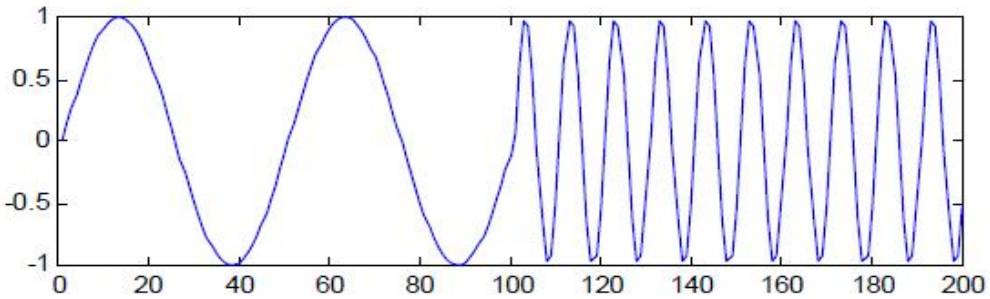

Figure 1. 8. Un signal avec des fréquences non constantes (non stationnaire)

- **Les sons quasi-stationnaires**

Un son est quasi-stationnaire s'il est stationnaire pour une période de temps particulière. Autrement dit, un son est dit quasi-stationnaire lorsque ses caractéristiques acoustiques telles que la fréquence et l'amplitude sont constantes pour une petite période puis varient dans le reste du temps. Un exemple sur un bruit quasi-stationnaire est le babillage constant d'une foule.

1.2.3.3. Différences clés entre les sons parole, musique et sons de l'environnement

Les sons environnementaux sont en général des sons non structurés, provenant de diverses sources. Contrairement à la parole et à la musique qui disposent d'un dictionnaire limité de phonèmes et de notes respectivement, les sons environnementaux qui sont très variés ne peuvent être décomposées en dessous unités bien significatives mais aussi ne peuvent suivre aucune règle ou grammaire prédéfinie [Alías et al., 2016]. En raison de la diversité inhérente, et leur nature non structurée il est difficile de trouver les meilleures caractéristiques qui décrivent

mieux les signaux audio. Le choix approprié de ces caractéristiques est essentiel dans la construction d'un système de reconnaissance robuste [Chu et al., 2009]. La parole utilise la structure phonétique, i.e., chaque morceau de parole peut être vu comme un ensemble de phonèmes qui peuvent être modélisés par des HMM, ce qui n'est pas le cas pour les signaux environnementaux. En revanche, Les sons environnementaux ne disposent pas de structure harmonique contrairement aux signaux musicaux qui présentent des modèles ou formes stationnaires telles que la mélodie et le rythme.

Deuxièmement, on peut généralement observer que la complexité du spectre des sons environnementaux est considérablement plus grande que celle des signaux de parole ou de musique, lorsque le signal est analysé dans le domaine fréquentiel [Alías et al., 2016]. En outre, on peut observer que les signaux de parole et de musique présentent généralement des structures harmoniques dans leurs spectres, un trait qui n'est pas si courant dans les sons environnementaux. En effet, Il convient de noter que les sons de parole et de musique sont tous deux composés d'un dictionnaire limité d'unités sonores : les phonèmes et les notes, respectivement. En revanche, la gamme des sons environnementaux est théoriquement infinie, puisque tout son présent dans l'environnement peut être inclus dans cette catégorie.

Finalement, La périodicité des sons environnementaux peut ne pas être aussi évidente comme c'est le cas pour la parole et la musique.

Le tableau 1.1 ci-dessous présente les principales différences entre les trois catégories de sons : musique parole et sons de l'environnement.

Tableau 1. 1. Comparaison entre les trois catégories de sons : musique, parole et sons de l'environnement

	Musique	**parole**	**Sons environnementaux**
structure	claire	Claire par rapport au langage	Structure non stable
source	Instruments ou humain	humain	Différentes sources
Unité	notes	Phonèmes	Non définie
Périodicité	périodique	périodique	Souvent non périodique
Type	Structuré (harmonie, rythme)	Structuré (phonétique)	Non structuré

En se basant sur ces différences clés entre les différents types de sons, les communautés de recherche proposent des solutions et techniques d'extraction de caractéristiques pour les sons environnementaux. Notons qu'il existe actuellement beaucoup de travaux dans ce domaine qui seront présentés plus en détails dans le chapitre suivant, dont quelques-uns utilisent les méthodes de reconnaissance et d'extraction de caractéristiques traditionnelles comme celles utilisées dans la parole et la musique et d'autres se basent sur les nouvelles solutions proposées et l'objectif étant toujours de comparer les résultats et d'en tirer des conclusions et par conséquent travailler plus pour les améliorer.

1.2.4. Taxonomie des sons

Les taxonomies permettent d'organiser et de structurer des concepts. Dans les domaines liés à l'audio, ils constituent la première étape vers la classification des sons en groupes basés sur différentes propriétés subjectives ou contextuelles [Schafer, 1993].

Des taxonomies disparates ont été développées sur la base d'une similarité subjective, d'une source sonore ou d'un contexte environnemental commun [Favory et al., 2018]. Cependant, comme les sons sont multimodaux, multiculturels et multiformes, il n'existe pas de taxonomie commune permettant d'organiser des collections de sons vastes et variées [Gerhard, 2003], [Favory et al., 2018].

Gerhard [Gerhard, 2003] par exemple, propose une taxonomie pour les sons de l'environnement en essayant de les mettre dans des catégories et sous catégories afin de faciliter leur traitement en se basant sur la perception humaine de ces sons. Certains travaux ont proposé des taxonomies pour les sons environnementaux, basées sur l'interaction des matériaux [Gaver, 1993] ou en fonction de leurs caractéristiques physiques [Schafer, 1993]. Comme expliqué aussi dans un travail récent de [Favory et al., 2018] qui porte sur la proposition de méthodes et outils pour faciliter l'annotation des sons dans de larges taxonomies : « Malgré tous les efforts fournis dans la conception de taxonomies spécifiques, la création de taxonomies plus grandes et plus générales a récemment retenu l'attention de la communauté des chercheurs » [Gemmeke et al., 2017], qui proposent l'ontologie 'AudioSet', l'une des plus grandes taxonomies qui structurent 632 catégories liées à l'audio.

Dans un contexte un peu spécifique qui est l'environnement maison, Sehili [Sehili, 2013] aussi propose une autre taxonomie adaptée à une application de télésurveillance des personnes âgées.

En effet, une taxonomie des sons de l'environnement n'est pas quelque chose de fixe mais elle peut varier d'une application à l'autre et l'élément responsable de cette variation est les **sons d'intérêt,** c'est à cause de ce dernier qu'on peut définir une taxonomie pour nos sons de l'environnement. Dans notre cas, par exemple, nous avons préféré d'utiliser le terme **sons de la vie courante** ou **sons de la vie quotidienne** pour désigner tout son de l'environnement qui peut se produire dans l'appartement y compris la parole et la musique.

La nature de l'application permet de préciser et limiter les sons d'intérêt et en considérant tout le reste des sons comme du bruit environnemental. Le son d'une machine à laver par exemple, peut être considéré comme du bruit pour une application de télésurveillance des personnes âgées, le son de toux peut être aussi considéré comme un bruit mais ceci peut être un son d'intérêt et important pour une application médicale afin de suivre l'état de santé de la personne surveillée. Plus de détails sont présentés dans le chapitre 3 qui décrit notre corpus de sons.

Nous considérons la taxonomie proposée par Gerhard [Gerhard, 2003] qui est une taxonomie standard et de référence et qui divise les sons en des catégories et sous catégories (voir figure 1.9).

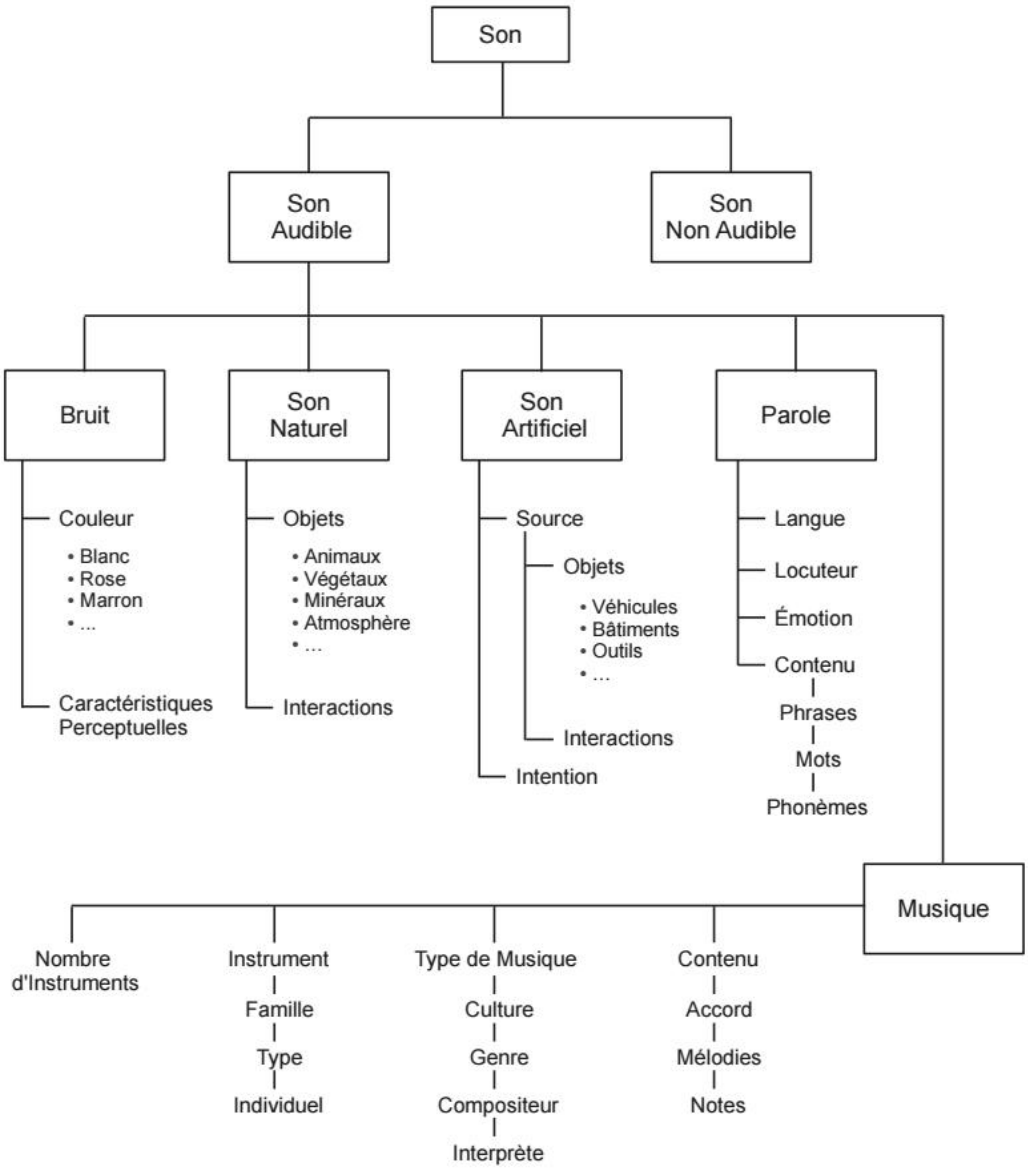

Figure 1. 9. Une taxonomie pour les sons perçus par les humains [Gerhard, 2003]

1.2.5. Processus de reconnaissance de son

Les problèmes de reconnaissance ou de classification, en général, passent par trois étapes principales notamment, le prétraitement des données, l'extraction des caractéristiques et la classification, tels que schématisé dans la figure 1.10.

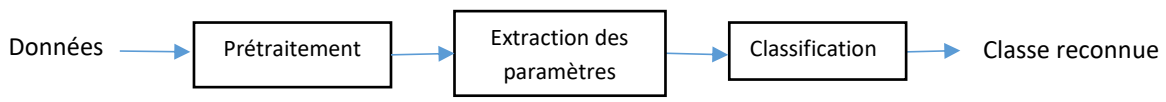

Figure 1. 10. Architecture générale d'un système de reconnaissance

Les systèmes de reconnaissance automatique de son ont le même principe que ceux des systèmes de reconnaissance automatique de la parole et de la reconnaissance automatique des formes. Les trois étapes clés pour l'implémentation d'un ASR sont, le prétraitement du signal, l'extraction des caractéristiques, et la classification.

En effet, Les systèmes de reconnaissance automatique de son ont le même principe que ceux des systèmes de reconnaissance automatique de la parole. La structure de base d'un système de classification des sons est illustrée dans la figure 1.11. Après l'extraction des vecteurs caractéristiques du signal, et après une phase d'apprentissage, une décision est prise sur la classe à laquelle appartient le signal. En effet, Il existe un grand nombre d'approches pour les classifieurs, dont beaucoup nécessitent beaucoup de puissance de calcul et / ou de mémoire. Dans ce qui suit, nous résumons les principales étapes.

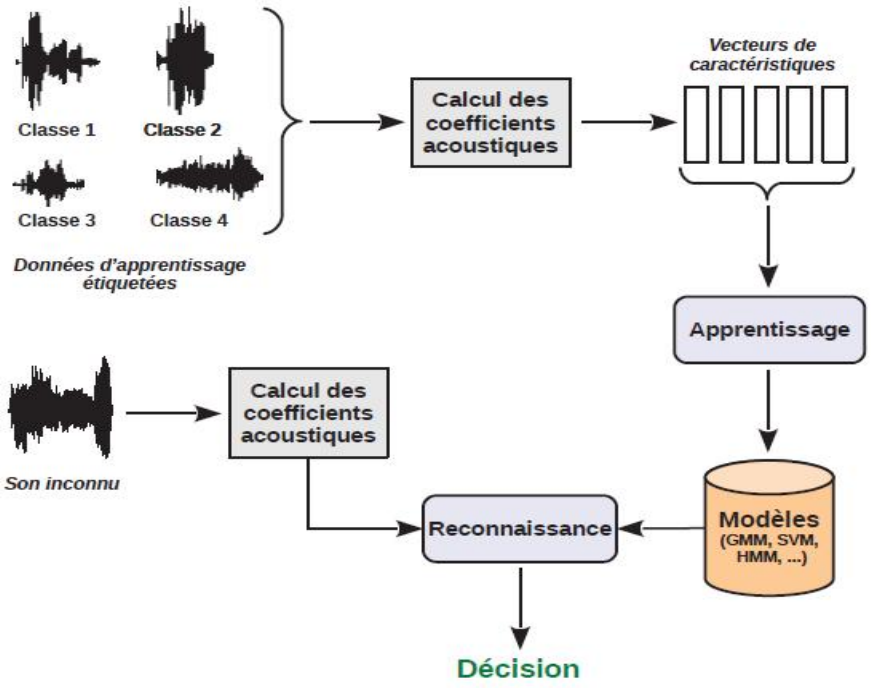

Figure 1. 11. Architecture d'un système de reconnaissance de son comme proposé dans [Sehili, 2013]

a) Prétraitement

D'après [Kefauver, 1999], le prétraitement des données pour la reconnaissance de son implique la prise du son de son environnement et le charger dans l'ordinateur via un microphone, ce qui signifie que le signal analogique produit par un microphone doit être converti en un format numérique via les techniques d'échantillonnage et de quantification.

Le prétraitement du signal consiste à préparer le signal sonore pour l'extraction de caractéristiques. Chachada, présente dans sa synthèse [Chachada et Kuo, 2013] les trois schémas de traitement des sons environnementaux. En règle générale, un signal est divisé en petites trames, souvent dans la plage de 10 à 30 ms ensuite, une fonction de fenêtre (window function) telle que la fenêtre de Hanning ou de Hamming est appliquée pour lisser le signal en vue d'une

analyse ultérieure. La fenêtre de Hamming semble être le choix préféré dans la plupart des systèmes ASR [Sharan et Moir, 2016]. Les systèmes de reconnaissance de la parole utilisent généralement une fréquence d'échantillonnage de 8000 Hz, tandis que les systèmes ASR utilisent une fréquence d'échantillonnage de 8 000 Hz ou plus, les valeurs courantes sont 16000Hz, 22050Hz et 44100Hz, qui dépendent de la bande de fréquences des signaux sonores pris en compte dans l'application.

Trois schémas de traitement des sons environnementaux sont couramment utilisés dans tout algorithme de reconnaissance des sons environnementaux [Chachada et Kuo, 2013] que nous résumons ci-dessous :

Traitement par trame (Framing-based processing en anglais) : dans cette technique les signaux audio sont divisés en trames en utilisant des fenêtres tels que la fenêtre de Hamming, puis les caractéristiques sont extraites à partir de chaque trame et l'ensemble de ces caractéristiques est utilisé comme une seule instance pour l'apprentissage et le test. La classification se fait pour chaque trame, c.-à-d., des trames consécutives peuvent appartenir à des classes différentes. L'inconvénient majeur de cette approche est la difficulté de sélectionner une longueur de fenêtre optimale convenant à toutes les classes ; certains événements sonores sont courts (par exemple, un coup de feu), tandis que d'autres sont plus longs (par exemple, le tonnerre). En effet, Si la longueur de la fenêtre est trop petite, les variations à long terme du signal ne seront pas capturées efficacement et les événements peuvent être divisés en plusieurs trames. Inversement, si la fenêtre est trop grande, il devient difficile de localiser les limites des segments entre les événements consécutifs, et plusieurs événements sonores peuvent tomber dans une seule trame.

Traitement basé sur les sous-trames : après division du signal en trames, chaque trame est segmentée en sous-trames, généralement avec un chevauchement, et les caractéristiques sont extraites de chaque sous-trame. Afin d'entrainer un classificateur, les caractéristiques extraites des sous-trames sont soit concaténées pour former un grand vecteur de caractéristiques, soit moyennées de manière à représenter une seule trame. Une autre possibilité consiste à entrainer un classificateur pour chaque sous-trame et à prendre une décision collective pour a trame sur la base des étiquettes de classe de toutes les sous-trames (par exemple, une règle de vote majoritaire). Ce modèle permet d'utiliser à la fois des caractéristiques non stationnaires et des classifieurs séquentiels. Même avec un classificateur non séquentiel, ce schéma de traitement permet de mieux représenter chaque trame, car la distribution collective de toutes les sous-trames permet de modéliser les caractéristiques intra-trame avec une plus grande précision. Cette méthode offre une plus grande souplesse dans la segmentation des événements sonores consécutifs sur la base des étiquettes de classe des sous-trames.

Traitement séquentiel : Les signaux audio sont divisés en petites unités appelées segments, d'une durée typique de 20 à 30 ms avec un chevauchement de 50 %. Le classifieur prend des décisions sur les étiquettes de classe et la segmentation en se basant sur les caractéristiques extraites de ces segments. Comparée aux deux méthodes mentionnées précédemment, cette

méthode permet de capturer la corrélation entre les segments et les variations à long terme du son environnemental sous-jacent. Cela peut être réalisé à l'aide d'un modèle de signal séquentiel, tel que les modèles de Markov cachés (HMM). Cette méthode peut être utilisée pour analyser des événements audio très courts.

Tout algorithme ESR suit fondamentalement l'un des trois schémas de traitement ci-dessus avec des variations mineures dans ses schémas de prétraitement et de sélection/réduction des caractéristiques.

b) *Extraction des caractéristiques*

L'extraction de caractéristiques est la première phase du système de reconnaissance des sons, elle consiste à extraire un ensemble de paramètres qui caractérisent le son et le définissent. L'extraction de caractéristiques est un processus qui consiste à effectuer une réduction efficace des données tout en préservant la quantité appropriée d'informations du signal [Agostini et al., 2001]. C'est la phase critique de tout système de reconnaissance, elle implique la sélection d'éléments des données d'entrée qui caractérisent ces informations de manière unique [Fukunaga, 1990]. L'étape d'extraction de caractéristique varie d'une application à une autre et dépend des classes de sons à reconnaitre. Il existe plusieurs techniques d'extraction de caractéristiques pour la reconnaissance de son qui sont divisées en deux grandes catégories : techniques d'extraction des caractéristiques stationnaires et techniques d'extraction des caractéristiques non stationnaires que nous décrivons par la suite. Cependant, quelle que soit la technique utilisée, les chercheurs affirment que l'extraction de caractéristiques est la partie la plus difficile du processus de reconnaissance [Dufaux, 2001].

En effet, **la phase d'extraction de caractéristiques** contient ou intègre les processus suivants [Dufaux, 2001] :

- ***Le prétraitement du signal*** : par suppression du bruit en utilisant par exemple le processus de filtrage.
- ***La projection du signal*** : en transformant le signal audio en entrée en un autre domaine où les paramètres du signal peuvent être facilement extraits. La transformée de Fourier est un exemple sur cette transformation qui permet une bonne discrimination des contenus fréquentiels.
- ***Extraction de caractéristiques :*** cette étape permet de calculer les caractéristiques désirées ou choisies tels que l'énergie, le nombre de passage par zéro (ZCR), les MFCC.
- ***Optimisation des caractéristiques :*** cette étape contient la normalisation des paramètres et la sélection des paramètres les plus représentatifs du signal. Cette étape sera expliquée plus en détail dans la section suivante.

c) *Apprentissage*

C'est une phase qui vient juste avant la phase de classification, elle dépend de l'algorithme de classification utilisé et s'applique sur les données dites d'apprentissage. A la fin de cette phase, on obtient des modèles représentatifs pour chaque classe, ces modèles présentent des modèles de référence utilisés par le classifieur pour décider de l'appartenance d'une donnée de test à l'un de ces modèles. Il existe différents protocoles pour le choix de la base de test et d'apprentissage [Istrate, 2003], nous citons :

- **Leave all in :** tous le corpus est utilisé pour l'apprentissage et en même temps pour le test.
- **La validation croisée :** une partie est utilisée pour l'apprentissage et l'autre pour le test. Parmi les protocoles de choix de ces parties d'apprentissage et de test il existe :
  - **Holdout Techniques :** les parties de test et d'apprentissage sont fixées au début.
  - **Leave one out :** utilise toutes les données sauf une pour le test (utilisé lorsque la taille du corpus est petite.
- **Ré-échantillonnage :** divise aléatoirement le corpus en une partie de test et d'apprentissage.

d) *Classification*

La classification est la troisième étape du processus de reconnaissance. Après la génération des vecteurs caractéristiques du signal en entrée dans la phase d'extraction de caractéristiques, une décision est prise sur la classe à laquelle appartient le signal. Le principe de la classification est le passage de l'espace de caractéristiques vers un espace de décision [Büchler et al., 2005]. Une classe correspondante est définie pour chaque point de l'espace de caractéristiques.

La classification consiste à prendre les caractéristiques générées à l'étape précédente et de relier chaque caractéristique à une forme particulière [Schalkoff, 1990]. Les frontières entre les classes se trouvent en effectuant une sorte d'apprentissage. Ceci est accompli avec un ensemble approprié de données sonores [Büchler et al., 2005]. Dans la classification, il est important de veiller à ce que les ensembles de test et les ensembles d'apprentissage soient enregistrés dans les mêmes conditions afin d'obtenir des résultats optimaux [Dufaux, 2001]. Cependant, Dans une analyse des techniques d'apprentissage et de test pour la reconnaissance de la parole, Murthy dans [Murthy et al., 1999] explique qu'il est nécessaire de collecter des données d'apprentissage dans divers environnements afin de garantir qu'un ensemble représentatif de données d'apprentissage est stocké dans la base de données.

Plusieurs techniques de classification peuvent être utilisées, notamment les modèles de Markov cachés, les réseaux de neurones, les GMM et les SVM. Toutes ces techniques utilisent le paradigme d'apprentissage et de test. L'apprentissage donne au système une série d'exemples d'un élément particulier afin qu'il puisse en apprendre les caractéristiques générales. Ensuite, lorsque le test est effectué, il peut identifier la classe de l'élément en cours de test [Dufaux, 2001].

Dans les sections qui suivent, nous présentons un aperçu sur les méthodes d'extraction de caractéristiques et de classification.

1.3. Méthodes d'extraction des caractéristiques

1.3.1. Taxonomie des méthodes d'extraction de caractéristiques audio

Alias et ses co-auteurs, dans [Alías et al., 2016], présentent une étude récente des méthodes d'extraction des caractéristiques pour les trois types de son : musique, parole et sons de l'environnement. Cette revue couvre les approches les plus élémentaires et classiques de l'extraction des paramètres audio, allant des années 1970 aux dernières contributions sur l'extraction des caractéristiques audio basées sur de nouveaux domaines du calcul et les paradigmes bio-inspirés. La principale référence de cette revue est l'étude de Mitrović [Mitrović et al., 2010] contenant les techniques classiques d'extraction des caractéristiques audio. Ensuite, l'auteur a étendu ces approches, en tenant compte des dernières avancées dans ce domaine de recherche, par des caractéristiques calculées dans le domaine temporel, fréquentiel et cepstral, puis, deux nouvelles techniques d'extraction de caractéristiques calculées dans le domaine des ondelettes et des images ont été décrites. Les techniques d'extraction de caractéristiques audio décrites sont classées en fonction de leur nature physique ou perceptuelle. La figure 1.12 ci-dessous présente la taxonomie des caractéristiques audio, et pour plus de détails sur les exemples de chacune de ces catégories se référencer à [Alías et al., 2016].

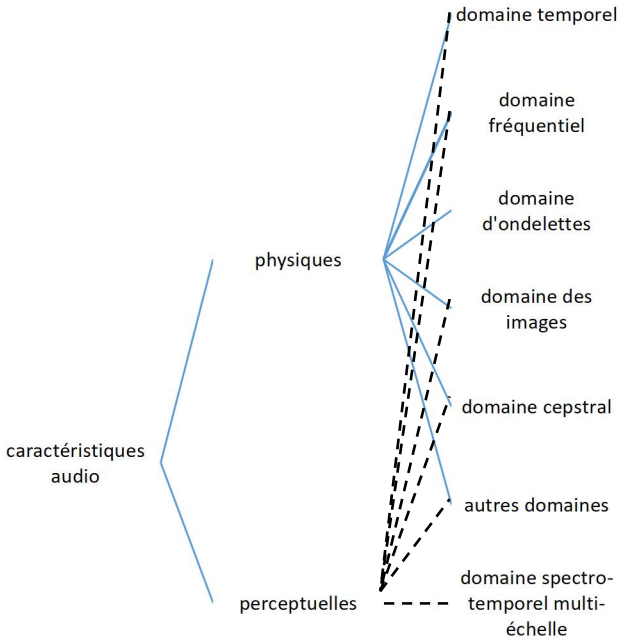

Figure 1. 12. Taxonomie des caractéristiques audio d'après [Alías et al., 2016]

L'étude de Chachada [Chachada et Kuo, 2013] aussi, est une étude approfondie et très intéressante sur les développements récents dans le domaine de la reconnaissance des sons de l'environnement où il présente la classification des méthodes RSE existantes en deux types principaux notamment, les techniques stationnaires et non stationnaires. En plus de ces

différentes études, il est intéressant de citer le travail de Crocco et ses co-auteurs [Crocco et al., 2016]. La taxonomie adoptée par [Crocco et al., 2016], subdivise les caractéristiques audio en six catégories : temporelle, spectrale, basée temps-fréquence, basée sur le cepstre, sur l'énergie, et sur la biologie ou la perception. En voici une description abrégée des caractéristiques calculées pour chaque classe :

Caractéristiques Temporelles : les caractéristiques temporelles sont directement extraites des échantillons de signal ou, plus généralement, d'une représentation temporelle du signal, nous citons : *ZCR, caractéristiques de la plage de fréquences du pitch, Waveform Minimum and Maximum.* En effet, les caractéristiques *Short-Time Average Zero-Crossing Rate, Logarithmic Short-Term Energy, Squared Short-Term Energy et Absolute Short-Term Energy* sont couramment utilisées pour identifier la parole dans un signal audio. Ces mêmes caractéristiques sont utilisées en combinaison avec d'autres paramètres acoustiques comme présenté dans [Delgado-Contreras et al., 2014a], pour la classification des environnements.

Caractéristiques spectrales (fréquentielles) : Les caractéristiques spectrales sont extraites du spectre du signal, comme *les coefficients de Fourier, BER (Band Energy Ratio), la largeur de bande*. Les caractéristiques spectrales sont utilisées avec succès pour la reconnaissance des instruments musicaux [Essid, 2005]. La plupart des caractéristiques spectrales donnent de faibles performances lorsqu'elles sont utilisées seules en raison de leur faible dimensionnalité, et en général, elles sont utilisées en combinaison avec des paramètres de plus grandes taille tels que les MFCC.

Caractéristiques temps-fréquence : Les caractéristiques temps-fréquence sont extraites d'une fonction de représentation bidimensionnelle du temps et de la fréquence, telle qu'un spectrogramme, ou une échelle de temps comme les ondelettes. Ces caractéristiques sont nombreuses, nous citons : *STFT, Spectrogramme, Coefficients des Ondelettes, coefficients de la transformée en ondelettes ou DWTC (Discrete Wavelet Transform Coefficients)* en anglais, fréquence fondamentale, harmonicité.

Caractéristiques à base de Cepstrum : Les caractéristiques à base de Cepstrum reposent sur le cepstrum ; une transformation non linéaire du spectre qui permet de représenter de manière compacte l'enveloppe du spectre, en ignorant les fines variations dans les intervalles de fréquences proches (par exemple, l'emplacement exact des harmoniques dans un signal périodique), nous citons : les *MFCC, MFCC et ses dérivatives,* Linear Prediction Cepstral Coefficients (*LPCC) et* Homomorphic Cepstral Coefficients (HCC)*.* Dans [Essid, 2005], l'auteur utilise les descripteurs cepstraux pour la classification des instruments musicaux et explique que la représentation cepstrale s'avère efficace pour de nombreuses tâches de classification audio telles que la discrimination parole/musique, la reconnaissance du genre ou encore la reconnaissance des instruments [Essid, 2005]. Il est à noter que d'après Chachada dans [Chachada et Kuo, 2013], les LPC et LPCC sont fréquemment utilisés pour la reconnaissance de la parole mais, ne sont pas utiles pour la RSE car ils incarnent le modèle source-filtre.

Les caractéristiques basées sur l'énergie : Ces caractéristiques méritent une classe distincte car leur calcul n'est généralement pas associé à une représentation de signal donnée (l'énergie peut être extraite d'un signal temporel, d'un spectre, d'un cepstre, etc.). Les fonctions basées sur l'énergie sont davantage impliquées dans les tâches d'extraction du contenu en premier plan, alors qu'elles ont tendance à être écartées pour la classification car, généralement, elles entraînent une augmentation de la variation intra-classe. Nous citons comme exemples : *l'énergie du signal, l'énergie de l'entropie.* Les caractéristiques basées sur l'énergie jouent un rôle mineur dans la classification des événements audio, principalement parce qu'elles sont sensibles à la distance qui les sépare de la source [Crocco et al., 2016], ce problème affecte évidemment leur classification. Plutôt que de cela, les caractéristiques basées sur la fréquence, en particulier celles qui capturent la forme du spectre, sont presque invariantes à la distance et, en même temps, très descriptives. Il en va de même pour les caractéristiques du cepstre.

Enfin, **les caractéristiques inspirées de la biologie ou la perception :** représentent une classe orthogonale aux précédentes : elles peuvent être fondées sur des représentations temporelles, spectrales, temps-fréquence ou cepstrales, mais partagent une inspiration commune issue de la psychophysiologie de l'appareil humain auditif et / ou vocal. En particulier, elles peuvent être subdivisées en trois sous-classes : (i) des caractéristiques imitant le traitement du système auditif humain, en particulier le filtrage cochléaire ; (ii) des caractéristiques reproduisant la perception psychologique des signaux auditifs ; et (iii) des caractéristiques construites en fonction du comportement physique du tract vocal humain, ce dernier étant limité à l'encodage des sons vocaux. Nous listons : Log Frequency Coefficients, Mel Frequency Coefficients, Spectral Features Based on Gammatone Filter Bank, Gammatone Cepstral Coefficients (GTCC), Linear Prediction Coefficients et ses dérivatives (LPC), Perceptual Linear Prediction Coefficients et Derivatives (PLP). Il convient de noter que les caractéristiques axées sur la perception suscitent un intérêt croissant ces dernières années [Crocco et al., 2016].

1.3.2. Méthodes d'extraction de caractéristiques stationnaires et non stationnaires

Les techniques d'extraction de caractéristiques peuvent être divisées en deux types principaux [Cowling et Sitte, 2003], [Cowling, 2004], [Chachada et Kuo, 2014] : l'extraction de caractéristiques stationnaires basée sur la fréquence et l'extraction de caractéristiques non-stationnaires basée sur le temps et la fréquence.

L'extraction de caractéristiques stationnaires *:* cette méthode a pour résultat un ensemble de fréquences du signal en entrée, mais sans fournir des informations sur l'emplacement de ces fréquences c.à.d. où ces fréquences apparaissent dans le signal. Les techniques ESR stationnaires sont dominées par les caractéristiques spectrales [Chachada et Kuo, 2013]. Bien que ces caractéristiques soient faciles à calculer, la modélisation des sons non stationnaires présente des limites [Cowling et Sitte, 2003]. Nous citons ici quelques techniques d'extraction de caractéristiques stationnaires : Homomorphic Cepstral Coefficients, MFCC, LPC, Mel Frequency

Linear Prediction Cepstral (LPC) Coefficients, Bark Frequency Cepstral Coefficients, Bark Frequency Linear Prediction Cepstral (LPC) Coefficients, Perceptual Linear Prediction (PLP).

L'extraction de caractéristiques non stationnaires : contrairement à la première méthode, elle divise le signal en des unités de temps ce qui permet d'identifier et de localiser les fréquences en précisant l'emplacement d'apparition. Cette technique nous permet de bien comprendre le signal [Cowling et Sitte, 2003]. Les techniques ESR non stationnaires utilisent des caractéristiques dérivées de la transformée en ondelettes, de la représentation fragmentée et du spectrogramme. Les méthodes basées sur les ondelettes donnent des résultats comparables aux méthodes stationnaires. La représentation fragmentée et les méthodes basées sur le spectrogramme sont généralement plus performantes. Cependant, Bien que les méthodes non stationnaires améliorent les performances, elles sont souvent coûteuses en ressources de calcul. En dernier lieu, selon Cowling [Cowling et Sitte, 2003], l'application de techniques d'extraction de caractéristiques stationnaires aux sons non-parole n'est pas idéale car la plupart des sons environnementaux sont par nature non stationnaires.

Toutes ces méthodes sont utilisées pour la reconnaissance de la parole mais aussi les deux méthodes Frequency extraction et Mel Frequency Cepstral Coefficients sont aussi utilisées pour la reconnaissance des instruments musicaux. Les principales techniques temps-fréquence qui sont communément mentionnées dans la littérature générale sont [Cowling, 2004] : Short-Time Fourier Transform (STFT), Fast (Discrete) Wavelet Transform (FWT), Continuous Wavelet Transform (CWT), Wigner-Ville Distribution (WVD).

1.3.3. Les MFCC

Dans cette partie, nous nous focalisons sur les paramètres acoustiques utilisés dans notre étude. Nous adoptons une présentation succincte indiquant brièvement la procédure de calcul.

Les MFCC (Mel-Frequency Cepstral Coefficients) appelés aussi coefficients cepstraux sur l'échelle de Mel sont des caractéristiques très utilisées dans le domaine de la reconnaissance de la parole et du locuteur. Le calcul des paramètres MFCC tient compte des particularités de l'oreille humaine et est basé sur la perception humaine de la parole [Istrate, 2003], [Davis et Mermelstein, 1980].

L'échelle de fréquence Mel (B) est définie par (1).

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

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

Les étapes suivies pour le calcul des MFCC sont schématisées dans la figure (figure 1.13) ci-dessous :

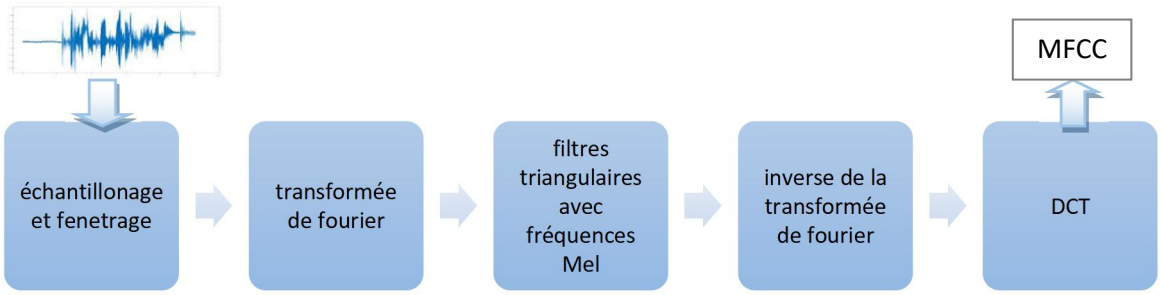

Figure 1. 13. Les étapes nécessaires pour le calcul des coefficients MFCC

Premièrement, le signal est décomposé en sections, puis une fenêtre de Hamming est appliquée. Ensuite, les fréquences Mel sont appliquées sur chaque segment fenêtré selon l'équation (2). L'application de ce filtre fournit une série de valeurs ; une pour chaque filtre. Une formule (formule 2) de coefficients cepstraux est appliquée sur cette série de valeurs pour enfin donner les MFCC, et ces caractéristiques sont ensuite collectées dans un seul vecteur caractéristique. En effet, la transformée de Fourier inverse est appliquée sur les coefficients en sortie des filtres puis une transformée en cosinus discrète (DCT ou Discrete Cosine Transform) est ensuite appliquée sur ces coefficients. Finalement nous gardons les coefficients de 2 à 13 et le reste sont éliminées.

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

Les Deltas et Delta-Deltas MFCC : appelés aussi coefficients différentiels et d'accélération respectivement. Ils présentent les trajectoires des MFCC à travers le temps. D'après des études précédentes, le calcul des trajectoires des MFCC et leur combinaison avec les MFCC augmente la performance d'un système de reconnaissance [Davis et Mermelstein, 1980], [Huang et al., 2001].

Les coefficients delta sont calculés par la formule suivante [Cowling, 2004]:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (3)$$

Où d_t est le coefficient delta, à partir d'un frame t calculé en fonction des coefficients statiques c_{t+N} jusqu'à c_{t-N}. De la même façon les coefficients delta deltas sont calculés mais en fonction des deltas coefficients mais pas à partir des coefficients statiques.

1.3.4. Traitements sur les caractéristiques

Nous citons dans cette section une série des traitements qui peuvent être appliqués sur les caractéristiques audio avant qu'ils soient fournis pour le classifieur à savoir la normalisation, la mise à l'échelle et la sélection des paramètres.

1.3.4.1. Normalisation des paramètres

La normalisation est une étape importante qui suit le processus d'extraction des caractéristiques afin d'éviter que l'amplitude typique d'une caractéristique soit d'un ou deux ordres de grandeur supérieure à celle d'une autre caractéristique. En effet, La normalisation n'est toujours nécessaire et ceci dépends des méthodes d'apprentissage utilisées. Lorsque la méthode est sensible à l'échelle telles que les méthodes basées sur la distance comme les SVM et les KNNs, la normalisation devient une étape nécessaire.

La normalisation est nécessaire car lorsqu'on forme les vecteurs de caractéristiques, et lorsque on les compare avec des distances standards, les caractéristiques d'une amplitude plus grande dominent, tandis que les autres n'ont pas une influence, ce qui explique la nécessité de normaliser les caractéristiques [Rabaoui et al., 2008]. De plus, normaliser les paramètres permet d'éviter un mauvais comportement numérique durant la phase de classification [Dufaux, 2001]. Comme a expliqué Dufaux dans sa thèse [Dufaux, 2001], le problème réside dans la possibilité de la non convergence de quelques algorithmes de classification récursifs lors de l'existence de petites valeurs numériques qui sont rencontrées dans le vecteur de caractéristiques lorsqu'il existe des différences importantes d'échelle entre les différentes dimensions de ce vecteur.

Avant l'apprentissage et le test par un SVM, les caractéristiques doivent être normalisées. En effet, il existe différentes méthodes de normalisation, Stolcke par exemple dans [Stolcke et al., 2007] utilise '*la normalisation par rang*' en remplaçant chaque caractéristique par son rang, ensuite mettre à l'échelle les rangs à une valeur entre 0 et 1. Cette méthode ne permet pas seulement de mettre à l'échelle les caractéristiques mais aussi de rendre leur distribution approximativement uniforme. Une alternative à la normalisation des caractéristiques dans le cas des SVM consiste à optimiser explicitement la fonction noyau pour minimiser l'erreur de classification. Une autre solution est d'utiliser les deux solutions à la fois c.à.d. normalisation des caractéristiques puis optimisation de la fonction noyau.

Pour récapituler, l'objectif de la normalisation est d'avoir des caractéristiques dans un rang spécifique de valeurs, afin d'éviter que le résultat du noyau linéaire (multiplication entre deux vecteurs dans le cas des SVM) soit « dominé » par certaines caractéristiques (celles aux rangs de valeurs larges) au détriment des autres (celles aux valeurs plus petites) [Dufaux, 2001], [Rabaoui et al., 2008], [Stolcke et al., 2007].

Delgado-Contreras et ses co-auteurs dans [Delgado-Contreras et al., 2014a] ont fait la normalisation des données afin de faciliter la manipulation du vecteur des données par le classifieur en utilisant deux équations :

$$NormZ = \frac{(x - mean(x))}{sd(x)} \quad (4)$$

afin d'obtenir un vecteur de caractéristiques dont la moyenne est zéro, et la deuxième équation (équation 5) pour mettre les données entre 0 et 1 :

$$PR = \frac{trunc(rank(x))}{length\ (x)} \quad (5)$$

La normalisation du signal du son se fait afin de limiter l'effet de conditions d'enregistrement variables sur les performances de classification [Essid, 2005].

1.3.4.2. Mise à l'échelle (Scaling)

La mise à l'échelle des données est une étape très importante, son principal avantage est d'éviter que les attributs aux plages numériques plus grandes dominent ceux des plages numériques plus petites. Un autre avantage est de faciliter les calculs. Dans le cas des SVM, la mise à l'échelle des attributs à l'intervalle [1 ; +1] ou [0 ;1] est recommandée [Hsu et al., 2003].

Une question que l'on peut poser est : quelle est la différence entre normalisation et mise à l'échelle ? Nous pouvons y répondre ainsi :

La mise à l'échelle appelée souvent *min-max scaling* est la transformation des données dans un rang spécifique. Tandis que la normalisation, c'est changer les observations de telle sorte qu'elles soient décrites comme une distribution normale. La normalisation est utile lorsqu'on envisage d'utiliser une forme quadratique telle que le produit scalaire ou un autre type de noyau pour quantifier la similarité entre n'importe quelle paire d'échantillons.

1.3.4.3. Traitement des sons de différentes durées

Les sons ont des tailles différentes, cependant pour pouvoir utiliser les SVM la taille du vecteur de caractéristiques doit être fixée. La méthode la plus adoptée dans les systèmes de reconnaissance de la parole est de supposer que les signaux sont composés d'un nombre fixe de sections. Rabaoui dans [Rabaoui et al., 2008] adopte cette méthode pour traiter les sons de différentes durées en partageant les trames d'un signal d'une durée T en 3 portions P1, P2 et P3, puis en construisant un vecteur caractéristiques composite de dimension 3n à partir de la moyenne des vecteurs caractéristiques des 3 régions, voir Figure 1.14.

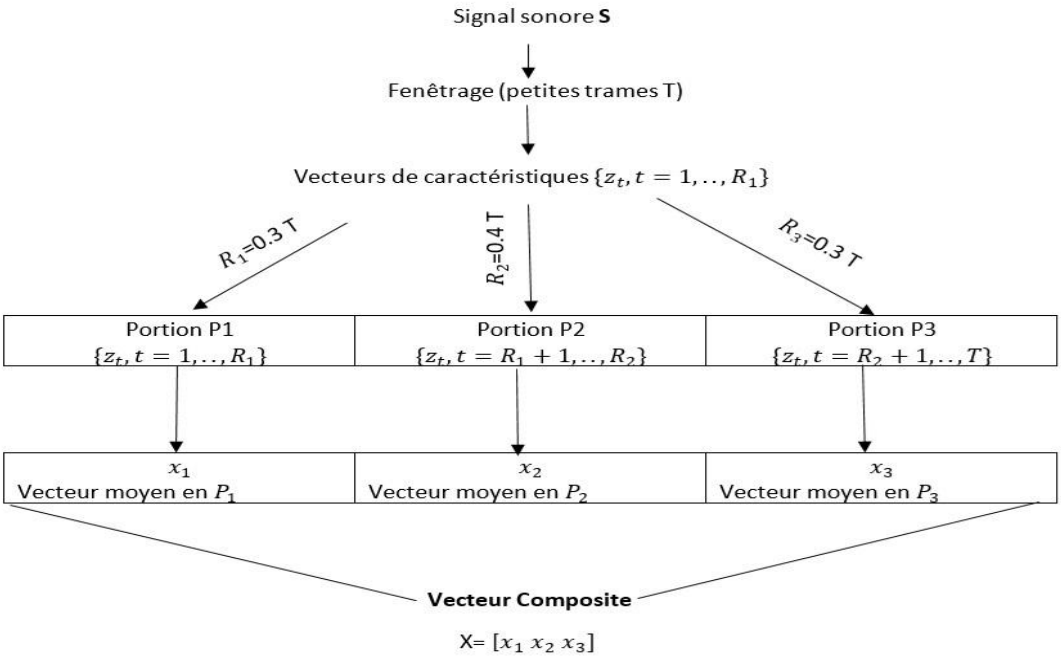

Figure 1. 14. Traitement des signaux de sons de différentes durées

1.3.4.4. Sélection des paramètres

Une fois les caractéristiques sont calculées, on obtient un vecteur de caractéristiques pour chaque son. Cette étape peut être précédée d'une étape de sélection des paramètres tels que l'Analyse en Composantes Principales (ACP) pour réduire le nombre des paramètres en enlevant par exemple les caractéristiques redondantes et celles qui n'apportent pas des informations pertinentes et celles qui ne sont pas discriminatives.

L'un des problèmes liés à l'utilisation d'un grand nombre de caractéristiques est qu'il existe de nombreuses caractéristiques potentiellement non pertinentes qui pourraient avoir un impact négatif sur la qualité de la classification. En utilisant des techniques de sélection des caractéristiques, nous pouvons choisir un ensemble de caractéristiques plus petit afin de réduire le coût de calcul et le temps d'exécution, ainsi que d'atteindre un taux de reconnaissance acceptable, voire supérieur. L'ajout de caractéristiques n'est pas toujours utile ; au fur et à mesure que la dimension des caractéristiques augmente, les points de données deviennent de plus en plus clairsemés [Chu et al., 2006]. Cela pose le problème de la sélection d'un sous-ensemble optimal de caractéristiques parmi un plus grand ensemble de caractéristiques qui produira le sous-ensemble le plus efficace. La solution optimale consiste à effectuer une recherche exhaustive de toutes les caractéristiques. Comme bien expliqué dans la thèse de [Dufaux, 2001], les meilleurs résultats de classification sont obtenus par l'utilisation d'un nombre limité des caractéristiques non corrélés et qui maximisent la séparabilité des classes.

Les méthodes de sélection des paramètres peuvent être classifiées de différentes manières cependant, la plus commune est la classification en filtres, enveloppement, intégrées et les méthodes hybrides [Jović et al., 2015]. Voici une description abrégée de ces méthodes :

Les méthodes filtres : elles sélectionnent les caractéristiques en se basant sur une mesure de performance sans tenir compte de l'algorithme de modélisation des données utilisé [Jović et al., 2015]. Elles utilisent les propriétés statistiques des caractéristiques pour éliminer les moins informatives en utilisant des algorithmes de classification. Cela se fait avant l'application des algorithmes de classification. A titre d'exemple de ces filtres, nous citons le critère de Fisher (Fisher Criterion Score (F)) qui calcule l'importance de chaque caractéristique indépendamment des autres caractéristiques en comparant sa corrélation vis-à-vis les étiquettes de sortie [Maldonado et al., 2009].

En réalité, Il existe plusieurs méthodes filtres, nous citons *CFS* (Correlation-based Feature Selection) [Hall, 1990], cette méthode consiste à sélectionner les sous-ensembles de caractéristiques et les classifie selon une fonction d'évaluation heuristique basée sur la corrélation. Consistancy-based [Dash et Liu, 2003], la méthode *mRMR* (minimum Redundancy Maximum Relevance) [Peng et al., 2005] qui sélectionne les caractéristiques les moins redondantes et qui ont le plus d'importance par rapport à la classe ciblée. Finalement, La méthode *reliefF* [Bolón-Canedo et al., 2013] ainsi que d'autres méthodes qui n'ont pas été citées.

Les méthodes d'enveloppement (wrappers) : cette approche consiste à mesurer l'utilité des caractéristiques en se basant sur les performances du classifieur. Elle utilise un modèle prédictif pour donner un score à un sous ensemble de caractéristiques [Maldonado et al., 2009]. Elle permet donc de sélectionner un sous-ensemble d'attributs qui permet d'atteindre les meilleures performances finales [Essid, 2005]. Cette approche donne des résultats plus précis que ceux obtenus par les méthodes filtres, mais elle est couteuse en temps de calcul [Maldonado et al., 2009]. La méthode wrapper nommée **WrapperSubsetEval** [Witten et Frank, 2005] en est un exemple, elle permet l'évaluation d'un ensemble d'apprentissage suivant un modèle d'apprentissage en utilisant la validation croisée pour estimer la précision pour un ensemble d'apprentissage. L'algorithme commence par un ensemble des caractéristiques qui est au départ vide ensuite il ajoute à chaque fois une caractéristique jusqu'à ce que les performances soient stables.

Les méthodes intégrées (embedded) : cette approche est très proche des « wrappers » elle permet l'optimisation du sous ensemble de caractéristiques et du classifieur en un seul processus [Essid, 2005]. Nous citons ici deux méthodes, notamment SVM-RFE et FS-P. SVM-RFE (Recursive Feature Elimination for Support Vector Machines) introduit dans [Guyon et al., 2002] par Guyon. Cette méthode indique l'élimination récursive des caractéristiques par utilisation d'un SVM en enlevant les caractéristiques les moins importantes indiquées par l'SVM. FS-P (Feature Selection—Perceptron) [Mejía-Lavalle et al., 2006] sélectionne les caractéristiques en se basant sur un réseau de neurones à propagation avant où les poids des interconnections sont des indicateurs sur les caractéristiques les plus pertinentes et ceci via l'apprentissage du classifieur dans le cas d'un apprentissage supervisé. Enfin, [Bolón-Canedo et al., 2013] est une bonne référence pour comprendre le principe de sélection des caractéristiques mais aussi faire son choix approprié. Les auteurs dans [Bolón-Canedo et al., 2013] présentent aussi une

comparaison entre les trois méthodes filtres, enveloppement et intégrées que nous pouvons résumer dans le tableau 1.2 ci-dessous :

Tableau 1. 2. Comparaison entre les trois méthodes de sélection des caractéristiques (filtres, wrappers et embedded)

Méthodes / Propriétés	Filtre	Embedded	Wrapper
Interaction avec le classifieur	indépendant	oui	oui
Temps de calcul	réduit	Inférieur à celui des wrappers	élevé
Capture des dépendances des caractéristiques	non	oui	Oui
Risque de surapprentissage	non	non	oui

Bilan sur la sélection des caractéristiques : nous avons présenté ci-dessus un aperçu sur les méthodes de sélection des caractéristiques et l'importance de cette étape dans la réduction de la complexité des systèmes d'apprentissage automatique et l'augmentation de ses performances en terme de précision et de taux de reconnaissance. Cependant, le choix d'une méthode de sélection des caractéristiques appropriée à un problème donné reste une question qui n'est pas assez facile à résoudre où on se trouve toujours dans un compromis performance et complexité en terme de puissance de calcul. D'après une étude comparative faite dans [Bolón-Canedo et al., 2013], les auteurs suggèrent d'utiliser les méthodes filtres, en particulier la méthode ***ReliefF***, car elles sont indépendantes des algorithmes d'apprentissage et sont plus rapides que les méthodes wrappers et embedded et avec une bonne capacité de généralisation. En conclusion, les méthodes ***filtres*** à nos connaissances et selon la nature de notre application qui nécessite un temps de réponse court sont les plus convenables et peuvent apporter des améliorations intéressantes si elles sont appliquées.

1.4. Méthodes de classification

Dans cette section, nous présentons un aperçu sur les méthodes de classification ainsi que les outils de classification que nous utilisons dans notre système de reconnaissance des sons, en expliquant leurs fondements théoriques.

1.4.1. Taxonomie des méthodes de classification

Une simple taxonomie pour les méthodes de classification a été présentée par [Crocco et al., 2016], qui divise les méthodes de classification en méthodes génératives et discriminatives.

Méthodes génératives : Dans ces méthodes, chaque classe d'événements audio a son propre classifieur, entraîné avec des échantillons de la même classe. Pour un échantillon de test, plusieurs classificateurs sont évalués (un pour chaque classe) et la probabilité à posteriori la plus élevée détermine le classificateur choisi et donc la classe choisie. Les modèles génératifs

GMM et HMM sont les plus répandus dans le domaine de la classification audio. En particulier, les HMM sont adaptés pour modéliser la variation temporelle du vecteur caractéristique sur des trames consécutives, ce qui permet une modélisation plus précise de chaque classe de son.

Méthodes discriminatives : D'autre part, les modèles discriminants tentent de construire directement le meilleur hyperplan de séparation dans l'espace des caractéristiques, en le divisant en sous-espaces et en séparant le plus grand nombre d'échantillons d'apprentissage pour chaque classe. Les réseaux de neurones artificiels (RNA) et les SVM sont les modèles de discrimination les plus répandus dans la tâche de classification audio.

Les points forts et complémentaires des deux modèles génératifs et discriminants peuvent être exploités par des stratégies hybrides.

Malgré qu'il existe plusieurs méthodes de classification, dans ce chapitre nous décrivons uniquement la méthode de classification basée sur les machines à vecteurs supports, sachant que cette dernière est la méthode utilisée dans notre expérimentation. Le choix de cette méthode est justifié par notre étude montrée dans le chapitre 2. La section qui suit montre une description détaillée de la méthode de classification basée sur les machines à vecteurs supports.

1.4.2. Les machines à vecteurs support (SVM)

L'SVM est considéré comme l'une des meilleures méthodes pour le traitement des problèmes de classification complexes [Uzkent et al., 2012], tels que la reconnaissance de la parole et la classification d'objets visuels. Dans son origine, l'SVM était un classifieur binaire, l'SVM est appelé aussi le classifieur de marge maximale [Uzkent et al., 2012], [Burges, 1998] à cause de sa puissance à trouver l'hyperplan de séparation optimal qui maximise la distance entre les points les plus proches des classes et l'hyperplan de séparation. A cause de la maximisation de la marge entre deux classes ceci donnera par conséquent de meilleures performances et surtout dans les espaces de grande dimension lorsqu'on utilise un nombre limité d'échantillons. Une des caractéristiques importantes des SVM est qu'ils sont des classifieurs peu sensibles à la dimension des vecteurs de descripteurs [Rabaoui, et al., 2007].

Il existe plusieurs implémentations logicielles des SVM. Cependant, l'outil LIBSVM [Hsu et al., 2010] développé en C++ reste l'outil le plus beau et efficace d'après plusieurs expérimentations, il existe aussi sous Matlab avec le nom OSU SVM [1] [Ma, et al., 2002]. Les références [Vert, 2001] et [Burges, 1998] présentent respectivement, une introduction et notions de base de l'SVM pour débutants, et un tutoriel plus détaillé. Un autre outil aussi intéressant est la bibliothèque Scikit-learn en Python qui est largement utilisée en apprentissage automatique, offrant une implémentation simple et efficace des SVM, et aussi la bibliothèque LIBSVM qui sont les deux utilisées pour la classification et la régression.

[1] The OSU SVM Support Vector Machine Toolbox for MATLAB uses the LIBSVM package.

1.4.2.1. Principe des SVM

a. L'SVM linéaire

Les SVM sont par essence des classificateurs bi-classes qui visent à séparer les exemples de chaque classe de manière à garder un maximum de marge de séparation entre n'importe quels exemples d'apprentissage, il s'agit de déterminer *l'hyperplan optimal.*

– **Cas des données séparables**

Soit à classifier les points suivants appartenant à deux classes différentes comme le montre la figure 1.15 ci-dessous.

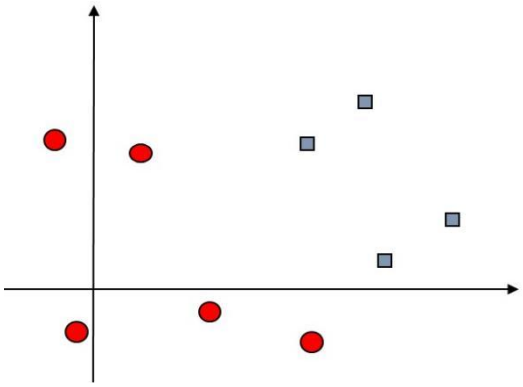

Figure 1. 15. Exemple des échantillons à classifier qui appartiennent à deux classes différentes

Le problème de la recherche de l'hyperplan optimal est un problème d'optimisation qui peut être résolu par des techniques d'optimisation tels que les multiplicateurs de Lagrange.

Plus formellement, étant donné un ensemble d'apprentissage où les données sont représentées ainsi : $\{\boldsymbol{x_i}, \boldsymbol{y_i}\}$, $\boldsymbol{i = 1, \ldots, l}, \boldsymbol{y_i} \in \{-1, 1\}, \boldsymbol{x_i} \in \boldsymbol{R^d}$. Les $\boldsymbol{x_i}$ sont les données ou les points d'apprentissage et les$\boldsymbol{y_i}$ sont les étiquettes.

Supposons qu'il y a des hyperplans qui séparent les exemples positifs des exemples négatifs. Les points appartenant aux hyperplans sont définis par

$$w.x + b = 0 \tag{6}$$

où $\boldsymbol{w}$ est la norme de l'hyperplan et b est le décalage par rapport à l'origine.

Il s'agit donc de séparer les exemples par un hyperplan optimal $H_{w0,b0}$ (figure 1.16.) de sorte à maximiser la marge entre les exemples d'apprentissage.

$$H_{w0,b0} = w_0.x + b_0 \tag{7}$$

En supposant que les données sont linéairement séparables, les exemples de chaque classe doivent satisfaire les conditions suivantes :

$$H_{w0,b0} = w.x_i + b \geq +1 \quad (8)$$

Pour $y_i = 1$, et

$$H_{w0,b0} = w.x_i + b \leq -1 \quad (9)$$

pour $y_i = -1$, ce qui peut être écrit par :

$$y_i(w.x_i + b) - 1 \geq 0, \forall i \quad (10)$$

Les deux hyperplans H_1, H_2 sont présentés respectivement par les formules (11) et (12):

$$H_1: w.x_i + b = +1 \quad (11)$$

$$H_2: w.x_i + b = -1 \quad (12)$$

La distance entre H_1, H_2 forme la marge et elle est $\frac{2}{\|w\|}$ car les deux hyperplans sont parallèles (la même normale w), et d'après (8) et (9) ils sont séparés et il n'existe aucun point commun entre les deux. La figure 1.16 en donne une illustration.

Le problème donc, s'exprime ainsi :

Maximiser la marge $\frac{2}{\|w\|}$ veut dire : minimiser $\|w\|$ et par conséquent, minimiser $\frac{1}{2}\|w\|^2$ sous les contraintes de l'équation (10).

En effet, Lorsqu'on veut trouver l'extremum d'une fonction avec contraintes, on utilise les *multiplicateurs de Lagrange*. On obtient alors :

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i (x_i.w + b) + \sum_{i=1}^{N} \alpha_i \quad (13)$$

Où les α_i sont les multiplicateurs de Lagrange.

Les points qui se trouvent sur les hyperplans $H1$ et $H2$ sont appelés les vecteurs supports tels qu'ils apparaissent dans la figure 1.16.

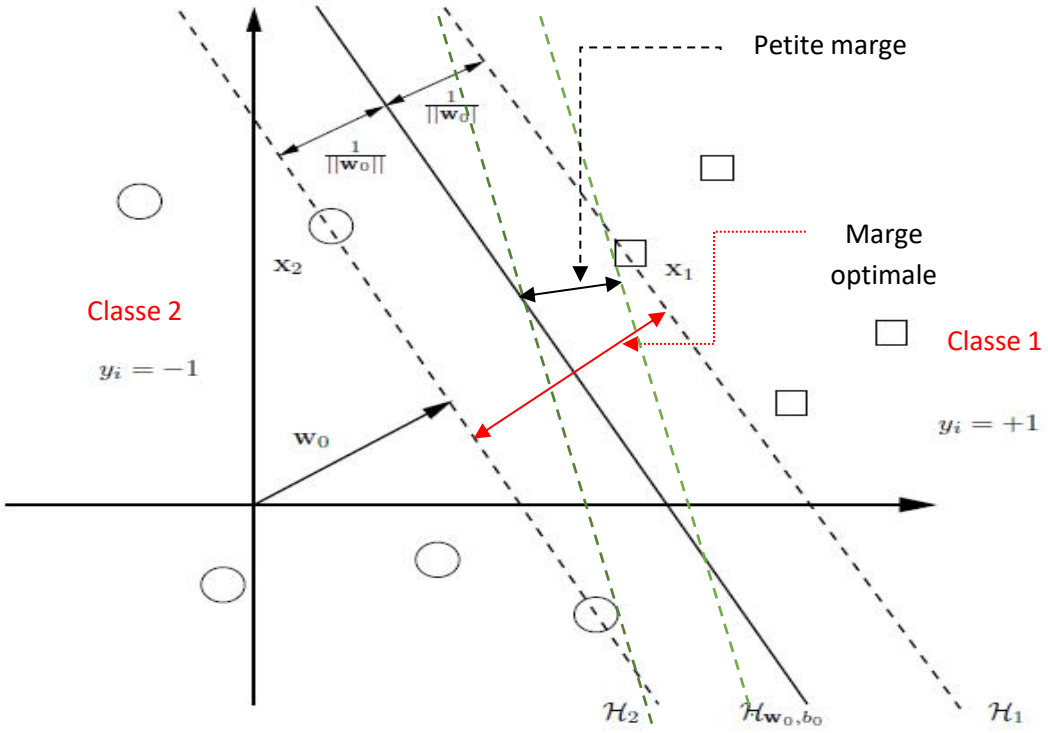

Figure 1. 16. Les hyperplans de séparation du nuage de points : un nombre infini des hyperplans de séparation mais un seul qui est optimale en maximisant la marge entre les points de deux classes

- **Cas des données non séparables**

Dans beaucoup de problèmes les données ne sont pas séparables d'où l'impossibilité de trouver un hyperplan de séparation entre les classes sans aucune erreur de classification. Une solution à ce problème consiste à rendre moins rigide les contraintes (10) en introduisant des variables ressort ou d'écart positives $\varepsilon_i \geq 0$ [Cortes et Vapnik, 1995], et la marge de séparation entre les classes devient souple dans ce cas. Le problème dans le cas des données non linéairement séparables devient :

Minimiser :

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\varepsilon_i \qquad (14)$$

avec les contraintes :

$$y_i(w.x_i + b) \geq 1 - \varepsilon_i, \forall i \qquad (25)$$

Où C>0 est un paramètre permettant de contrôler le compromis entre le fait de maximiser la marge et minimiser les erreurs de classification commises sur l'ensemble d'apprentissage.

Les variables ressort ε_i mesurent l'écart d'un point de données par rapport à la condition idéale de séparabilité du modèle:

- Pour $0 \leq \varepsilon_i \leq 1$, le point de données se situe dans la région de séparation mais sur le côté droit de la surface de décision (classification correcte).
- Pour $\varepsilon_i > 1$, le point de données tombe du mauvais côté de l'hyperplan de séparation (erreur de classification).

b. L'SVM non linéaire

L'idée de l'SVM non linéaire est de garantir une séparation linéaire pour des formes non linéairement séparables en mappant les données vers un espace dimensionnel supérieur en utilisant une fonction de transformation.

Remarquons tout d'abord que les données apparaissent dans le problème d'apprentissage sous la forme de produits scalaires $x_i \cdot x_j$. Supposons maintenant que nous avons d'abord mappé les données sur un autre espace euclidien H (éventuellement de dimension infinie), en utilisant une fonction de mapping que nous appelons: φ

φ: $R^d \rightarrow H$

Dans ce cas, l'algorithme d'apprentissage va dépendre uniquement des produits scalaires c.à.d. des fonctions de la forme $\varphi(x_i).\varphi(x_j)$.

S'il existe une *fonction noyau K* telle que

$$K(x_i, x_j) = \varphi(x_i).\varphi(x_j), \tag{16}$$

La fonction finale de décision est définie par :

$$f(x) = \sum_{i=1}^{n_s} \alpha_i y_i k(s_i, x) + b \tag{3}$$

Où n_s est le nombre de vecteurs de support et les s_i sont les vecteurs de support.

Alors, nous n'avons pas besoin de connaître ou de calculer φ. En d'autres termes, la fonction du noyau définit les produits internes dans l'espace transformé. La solution donc, réside dans l'utilisation de noyaux autres que le noyau linéaire. Citons-nous ici les noyaux les plus utilisés :

- Noyau linéaire

$$k(x, y) = x.y \tag{18}$$

- Noyau gaussien ou RBF (Radial Basis Function)

$$e^{-\|x-y\|^2/2\sigma^2} \tag{19}$$

- Noyau polynômial

$$(x.y + 1)^p \tag{20}$$

- Noyau sigmoïde

$$k(x, y) = tanh\ (\alpha x.\ y + c) \tag{21}$$

1.4.2.2. Les SVM dans le cas de plusieurs classes

L'SVM est de nature un classifieur binaire et afin de traiter le cas d'un SVM multi-classes il existe deux solutions possibles : soit plusieurs classifieurs binaires doivent être construits, soit un problème d'optimisation plus important est nécessaire où le nombre de variables est proportionnel au nombre de classes. Cependant, il est généralement très coûteux en termes de puissance de calcul de résoudre un problème multi-classe que de résoudre un problème binaire [Hsu et Lin, 2002].

La première solution est la plus adoptée, pour cela, plusieurs SVM binaires sont construits puis combinés en se basant sur quelques stratégies telles que : ***un-contre-un*** *(en anglais* one-against-one (OAO)), ***un-contre-tous*** (one-against-the-rest (OAR)), et **graphe orienté acyclique** (en anglais *directed acyclic graph* ou **DAG**), les trois premières stratégies ont été testées et comparées par [Uzkent et al., 2012] dans un travail de reconnaissance de quelques types de son (7 classes) en utilisant les SVM. Voici une description abrégée des trois méthodes :

Supposant qu'on a N classes, il s'agit donc d'un problème de classification N-classes.

- Pour la méthode **un-contre-tous** (OAR), les N classifieurs binaires sont entrainés de telle sorte que chaque classifieur sépare une classe des N-1 classes restantes. Tous les classifieurs sont entrainés sur toute la base d'apprentissage, et pour un échantillon de la base de test l'étiquette de la classe est déterminée par sélection du classifieur qui fournit la plus grande valeur de la fonction de décision en sortie.
- Concernant la méthode **OAO,** N(N-1)/2 classifieurs binaires sont construits, la décision est prise par utilisation de l'algorithme ***max-wins*** qui peut être expliqué ainsi : chaque classifieur OAO fait un vote pour sa classe préférée, et la décision finale est faite pour la classe avec le plus de votes.
- Finalement, pour la stratégie **DAG**, N(N-1)/2 classifieurs binaires sont entrainés avec utilisation d'un graphe orienté acyclique (DAG) durant la phase de test.

Il existe également d'autres approches telles que **les codes correcteurs d'erreurs** (en englais **Error Correcting Output Code** (ECOC)**,** c'est une méta-méthode qui combine de nombreux classifieurs binaires afin de résoudre le problème multi-classe [Dietterich et Bakiri, 1995]. Selon une expérimentation faite par Hsu et Lin dans [Hsu et Lin, 2002], les deux méthodes *un-contre-un* et *DAG* semblent les plus pratiques vu leur temps de calcul réduit par rapport à la méthode *un-contre-tous*.

1.5. CONCLUSION

Dans ce chapitre nous avons abordé les notions importantes du domaine et les concepts clés liés à la reconnaissance de sons. Nous avons adopté une approche ascendante dans la rédaction de la première partie de ce chapitre en commençant par l'élément le plus fondamental dans cette recherche qui est 'le son' pour ensuite aller progressivement vers des concepts plus larges et plus complexes, ceci nous a permis de mieux faciliter la lecture et la compréhension de ce chapitre. Une deuxième partie est consacrée pour les méthodes d'extraction de caractéristiques suivie d'une dernière partie qui est les méthodes de classification. Dans chacune de ces deux dernières parties nous avons présenté les méthodes existantes sous forme de taxonomies. Dans la section méthodes d'extraction de caractéristique audio, une bonne partie est consacrée pour les coefficients MFCC vus qu'ils utilisés dans ce travail. De même la section méthodes de classification, nous avons mis l'accent sur la présentation de la méthode utilisée et choisie dans notre travail pour des raisons qui sont bien expliquées dans le chapitre suivant.

En effet, plusieurs conclusions peuvent être tirées de ce chapitre que nous pouvons résumer ainsi : Les sons environnementaux sont très variés et diversifiés

- La plupart des sons environnementaux sont non structurés et contrairement à la parole et la musique, ils ne présentent pas des traits spécifiques et ils sont en général non stationnaires et impulsifs. Ces caractéristiques rendent la classification des sons une tâche plus compliquée que ce soit au niveau du choix des meilleurs paramètres acoustiques qui offrent une présentation fidèle des sons qui différent d'un son à un autre, ou des méthodes de classification qui sont de nature utilisées pour la classification de la parole et la musique.
- Il existe un nombre important de méthodes d'extraction de caractéristiques audio et nous avons présenté une taxonomie qui les organise en classes.
- Les méthodes de classification sont aussi nombreuses et l'accent a été mis sur la méthode SVM.

De notre point de vue, ce présent chapitre est d'une grande importance et beaucoup d'efforts lui sont consacrés. D'une part, la richesse du domaine et le nombre important de concepts clés que nous devons présenter, d'autre part, se limiter au notions les plus nécessaires et pertinentes pour ne pas trop encombrer le chapitre et par conséquent s'éloigner des informations principales.

Une question importante que nous pouvons poser après la fin de ce chapitre est la suivante : où sont appliquées toutes ces méthodes ? Peuvent-elles être appliquées dans le domaine de la reconnaissance des sons de l'environnement ? Existent-ils des travaux de reconnaissance des sons qui l'utilisaient ? C'est ce que nous essayons d'expliquer dans le chapitre suivant.

CHAPITRE 2
Travaux et méthodes

Dans ce chapitre, nous décrivons les travaux et les méthodes utilisées pour la reconnaissance des sons de l'environnement. Ensuite, nous abordons la reconnaissance des sons dans un contexte domotique et en particulier pour la détection des situations de détresse dans un habitat. L'idée de base de ce chapitre est de donner un état de l'art sur le domaine de reconnaissance des sons et les solutions proposées.

2.1. Introduction

La reconnaissance de sons est un domaine récent par rapport à la reconnaissance de la parole et la musique et elle n'a pas encore ses propres méthodes. Ses domaines d'application sont très variés, tels que la surveillance audio, la télé-santé, et la détection des événements multimédia. En effet, dans la vie courante, nous rencontrons une large variété de sons tels que les sons de pas, les cris, et le tonnerre. Les événements sonores se produisent souvent dans des environnements non structurés de la vie réelle. Des facteurs tels que le bruit ambiant et les sources qui se chevauchent sont présents dans les environnements non structurés et ils peuvent introduire un degré élevé de variation entre les événements sonores de la même classe [Cakir et al., 2017].

Cependant, chaque application s'intéresse à un ou plusieurs sons selon l'objectif visé, ce qui a rendu difficile la comparaison des différents systèmes proposés. Par conséquent, pour mieux clarifier les divers choix des applications en ce qui concerne les méthodes de classification et les méthodes d'extraction de caractéristiques, et en vue de la recherche des méthodes convenables pour la RSE, nous allons traiter ce chapitre de la façon suivante : en premier lieu, nous abordons les différents travaux réalisés dans le domaine de la reconnaissance de sons, ensuite nous présentons les différences clés entre la reconnaissance de la parole et de la musique pour ensuite présenter les méthodes utilisés pour la reconnaissance de sons en étudiant et comparant différents travaux sur la reconnaissance de sons. Une autre section est consacrée pour une synthèse des travaux sur les systèmes de de reconnaissance des sons basés sur les méthodes d'apprentissage profond dans l'objet de comparaison avec les méthodes de classification traditionnelles. Enfin de ce chapitre, nous présentons et synthétisons des travaux sur la détection des situations de détresse en utilisant le canal audio où l'objectif étant toujours d'analyser et de présenter les méthodes utilisées, sachant qu'un critère important dans ces synthèses est le nombre de classes à reconnaître et les sons ciblés.

2.2. Reconnaissance de son

2.2.1. Domaine de la reconnaissance de son

Tout environnement donné contient généralement un certain nombre de sons différents. Dans la littérature ancienne, ces sons étaient souvent divisés en parole et en non parole. La tâche de la classification des sons non parole est maintenant plus communément appelée reconnaissance automatique du son (ASR). Il est également appelé reconnaissance d'événements sonores (SER) et détection d'événements acoustiques dans certains contextes [Sharan et Moir, 2016].

Un système ASR vise à reconnaître automatiquement les sons à l'aide de techniques de traitement du signal et d'apprentissage automatique. En principe, il est très similaire à un système de reconnaissance automatique de la parole, à la différence que le signal d'entrée est non vocal. Alors que les recherches sur la reconnaissance de la parole ont fait l'objet d'une

attention considérable au cours des dernières décennies, celles portant sur les ASR ne semblent que s'être intensifiées au cours des deux dernières décennies environ [Sharan et Moir, 2016].

L'analyse de scène auditive (ASA ou Auditory Scence Analysis) : Lorsque nous parlons de la reconnaissance ou en particulier la classification de sons, nous désignons la classification des données en entrée après extraction des caractéristiques et dans ce cas on s'intéresse au résultat de la classification. Cependant, lorsque les données ne sont pas séparées, le cas d'un environnement réel, où les sons sont chevauchés, il est nécessaire de les séparer d'abord pour pouvoir ensuite les classifier. Tout ce traitement désigne *L'analyse de scène auditive* qui est une extension pour la classification de son.

L'ASA est le processus par lequel le système auditif sépare les sons individuels dans des situations de monde naturel, dans lequel ces sons sont généralement imbriqués et superposés dans le temps et leurs composants entrelacés et imbriqués en fréquence [Bregman, 1990]. Elle décrit des mécanismes et des stratégies de traitement sur lesquels le système auditif s'appuie dans l'analyse de l'environnement acoustique. Ce processus implique l'analyse d'une scène, tel que le bruit produit lors d'un cocktail de sons, puis la séparation et la classification des sons dans l'environnement. En effet, L'ASA désigne le processus qui consiste à utiliser les caractéristiques des événements acoustiques d'un environnement donné, qui ont lieu dans un certain laps de temps, en vue de les reconnaître [Bregman, 1990] et [Sehili, 2013].

Le ***Cocktail Party*** présente le vrai problème de l'ASA [Bregman, 1990], qui désigne la présence de plusieurs sources sonores dans un environnement donnant par conséquent une superposition de sons, et le problème réside donc dans la difficulté à pouvoir séparer ces sons et à interpréter chaque son à part. Ou bien pouvoir se préoccuper d'un seul type de son et négliger ou éliminer les autres, comme c'est le cas pour un être humain qui peut se concentrer sur une conversation qui se trouve dans une multitude d'autres sources sans être gêné par les autres sons qui sont produits dans son environnement et c'est grâce au système nerveux humain.

L'analyse de scène auditive par ordinateur (Computational Auditory Scene Analysis (CASA) : la question posée est : Comment peut-on concevoir des systèmes « d'écoute machine » capables d'écouter des « cocktails » ? Les auditeurs sont capables de séparer de manière perceptible une source sonore d'un mélange acoustique, telle qu'une seule voix d'un mélange d'autres voix et de la musique lors d'un cocktail animé [Rosenthal et Okuno, 1998]. La CASA a pris ses principes à partir du domaine de l'ASA qui est décrit par Bregman dans son livre publié en 1990 [Bregman, 1990] où il a établi une analogie entre la perception des scènes auditives et des scènes visuelles et décrit un framework pour comprendre l'organisation perceptuelle du son.

La figure 2.1 montre la relation entre l'analyse de la scène auditive et la reconnaissance de son.

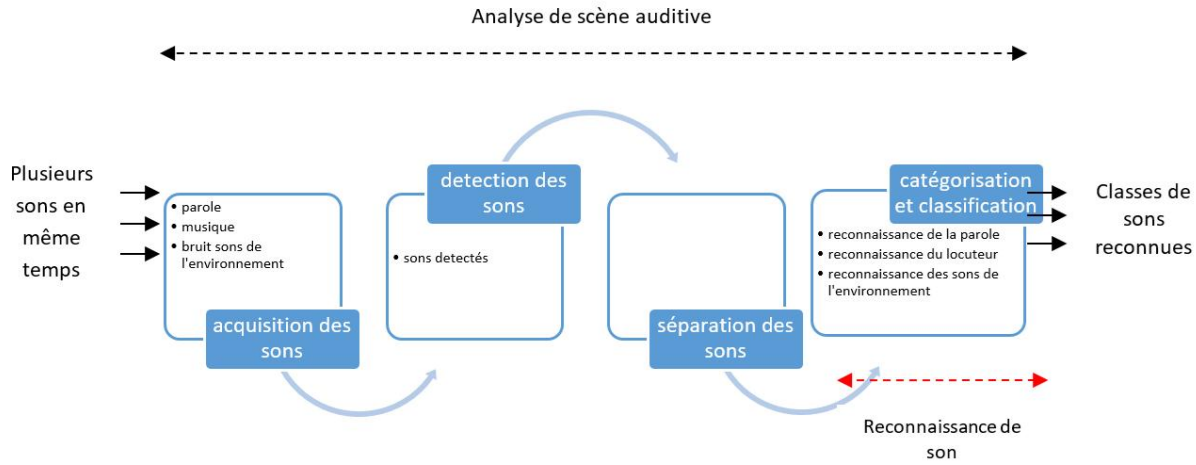

Figure 2. 1. Analyse de scène auditive et reconnaissance de son

2.2.2. Domaine d'application

La reconnaissance des sons environnementaux ou RSE est un domaine d'application très important pour les systèmes de reconnaissance automatique des sons appelés *Automatic Sound Recognition* (ASR) qui vise à reconnaître les sons automatiquement à l'aide des techniques de traitement du signal et d'apprentissage automatique [Sharan et Moir, 2016]. Comme décrit dans [Chachada et Kuo, 2013] et [Sharan et Moir, 2016], la RSE peut être appliquée dans plusieurs domaines. La RSE peut être utile pour les applications de recherche de sons similairement à la recherche d'image et de vidéo par contenu, comme elle peut être aussi appliquée dans l'étiquetage automatique des fichiers audio avec des descripteurs pour l'extraction de l'audio par mots clés. Un deuxième domaine d'application est dans les robots : il peut être intégré dans un système de navigation des robots en complément avec la fonction de vision qui n'est pas toujours suffisante car elle est limitée par le champ visuel [You et Li, 2012]. Un autre domaine d'application est les maisons intelligentes ou dans les habitats pour la surveillance des personnes âgées ou handicapées vivant seules dans leurs propres maisons en combinaison avec l'analyse de l'image et de la vidéo [Chachada et Kuo, 2013]. Dans les maisons intelligentes la RSE peut être utilisée pour de nombreuses applications à l'intérieur de la maison, telles que la quantification de la consommation d'eau [Ibarz et al., 2008], les bris de fenêtres pour des fins de sécurité, aboiement de chiens, cris des bébés, etc. Pour une meilleure consultation des différentes applications de RSE, je trouve le chapitre [Krstulović, 2018] dur livre [Plumbley et al., 2018] une source très intéressante, et de même pour le livre qui englobe les notions de base de l'analyse des scènes audio.

Du point de vue de [Dufaux, 2001], l'utilisation des RSE se concentre sur deux domaines principaux qui sont l'identification et la détection automatique d'alarmes pour les systèmes de surveillance et l'assistance et l'aide à l'autonomie des personnes âgées ayant des

problèmes auditifs. De même que pour [Dufaux, 2001], Cowling dans thèse [Cowling, 2004] a affirmé que les domaines tels que la technologie d'aide auditive et les systèmes de sécurité peuvent bénéficier de la recherche dans un système qui permet d'identifier les sons environnementaux autres que la parole. Les domaines d'applications des RSE ainsi que les travaux réalisés sont bien détaillés dans [Chachada et Kuo, 2013], [Sharan et Moir, 2016].

2.3. Des méthodes pour la reconnaissance de son ?

2.3.1. Principales approches pour la RSE

Une autre question importante qu'on peut poser dans cette thèse et est-ce que la RSE a ses propres méthodes ? ou bien elle utilise des méthodes issues des autres domaines de reconnaissance tels que la reconnaissance de la parole et la reconnaissance de locuteur. C'est à cette question que nous essayons de répondre dans cette section.

La synthèse effectuée dans [Sehili, 2013] a comme objectif de mettre les différents travaux dans le domaine de la reconnaissance des sons dans des catégories en mettant l'accent sur les méthodes adoptées. Ces catégories en fait, représentent les quatre approches utilisées pour la reconnaissance de son : la parole, le locuteur, le système auditif humain, et les techniques de traitement d'image.

La reconnaissance des événements acoustiques est un domaine qui est récent par rapport à la reconnaissance automatique de la parole, la reconnaissance automatique du locuteur et la transcription musicale, et à l'heure actuelle, ce domaine n'a pas trouvé encore ses propres méthodes. La REA se base sur les méthodes ou approches fondées sur la reconnaissance de la parole, la reconnaissance du locuteur, le système auditif humain et les techniques de traitement d'image via la transformation de l'information audio en une image à deux dimensions où le problème devient un problème de reconnaissance des images. Sehili dans [Sehili, 2013], a fait une étude sur les méthodes de reconnaissance et d'extraction des caractéristiques en comparant les systèmes proposés en vue de la recherche des méthodes les plus appropriées pour le domaine de REA. A partir de ces quatre catégories, une analyse a été faite pour pouvoir comparer les travaux issus de chacune de ces catégories et discuter les résultats.

Partant de la synthèse effectuée dans [Sehili, 2013], et en utilisant d'autres références et travaux, nous pouvons résumer les approches utilisées en reconnaissance de sons dans ce qui suit.

Approches fondées sur la reconnaissance de la parole : Les MFCC avec les HMM est l'approche la plus utilisée en RAP (Reconnaissance Automatique de la Parole) [Sehili, 2013],[Baker et al., 2009]. En effet, Les MFCC seuls ne peuvent pas garder l'information temporelle du signal. Pour cette raison, il existe plusieurs techniques qui peuvent être utilisées pour inclure l'information temporelle tels que les coefficients delta et double-delta et la technique RASTA (RelAtive SpecTrAl). D'après Cowling [Cowling et Sitte, 2002], les HMM

ne sont pas appropriés pour l'analyse des sons de l'environnement en raison de l'absence d'un alphabet phonétique, comme le cas de la parole. Cependant, plusieurs travaux existent et utilisent les HMM.

Tel qu'expliqué par Chachada dans [Chachada et Kuo, 2013], en général, les sons environnementaux comme les sons de tonnerre et de tempête n'ont aucune sous structure claire telle que les phonèmes ce qui rend leur modélisation avec des HMM une tâche difficile. De même, en comparaison avec les signaux musicaux, les sons de l'environnement ne présentent aucune structure ou forme stationnaire significative telle que la mélodie et le rythme.

En plus des HMM, il existe d'autres méthodes telles que les réseaux de neurones artificiels (ANN), les GMM, la quantification vectorielle (VQ pour Vector Quantization) et la déformation temporelle dynamique (DTW pour Dynamic Time Warping).

Approches fondées sur la reconnaissance du locuteur **:** Les méthodes de classification utilisées en RAL sont largement utilisées en REA. Nous citons dans cette partie quelques méthodes ainsi que les travaux l'utilisant pour la REA.

A nos connaissances, le premier travail qui a fait une comparaison entre les techniques de reconnaissance des sons environnementaux est celui de Cowling et Sitte [Cowling et Sitte, 2003], vient ensuite, le papier de Chachada [Chachada et Kuo, 2013] après une dizaine d'années qui s'agit d'une étude sur les nouveautés sur le domaine de RSE en terme des techniques d'extraction des caractéristiques et des techniques de classification. Parmi les conclusions importantes de sa revue est que les méthodes ESR stationnaires sont faciles à calculer mais il existe des limitations dans la modélisation des sons non stationnaires. Contrairement aux techniques non stationnaires, malgré qu'elles donnent de meilleures performances mais elles sont souvent coûteuses en temps de calcul.

Parmi les techniques de classification les plus utilisées dans le domaine de la reconnaissance de la parole/du locuteur, comme cité dans [Istrate, 2003], nous avons : *Les HMM, Les GMM, L'alignement temporel dynamique (DTW - Dynamic Time Warping) et Les SVM.*

Approches fondées sur le système auditif humain **:** Il existe un ensemble important de travaux qui modélisent le fonctionnement de la cochlée du système auditif humain en proposant un ensemble de filtres auditifs. Les filtres gammatones [Patterson et al., 1987] [Patterson et al., 1995] sont considérés comme l'une des modélisations de la perception les plus connues.

Différents travaux sur la reconnaissance de la parole et du locuteur et même de la reconnaissance de son utilisent les filtres auditifs. Nous citons à titre d'exemple le travail de [Valero et Alías, 2012] qui utilise les filtres gammatones pour la reconnaissance de sons de l'environnement. Un autre travail [Agrawal et al., 2017] utilise une nouvelle forme de filtres

gammatones en les combinant avec le Teager Energy Operator (TEO) pour la classification des sons de l'environnement. Dans un autre travail [Park et Yoo, 2020], les auteurs utilisent les bancs de filtres gammatones pour la classification des sons de l'environnement.

Qi et ses co-auteurs [Qi et al., 2013] utilisent les caractéristiques basées sur les filtres gammatones pour la reconnaissance de la parole. L'idée du travail présenté dans [Marković et al., 2017] est l'utilisation des coefficients cepstraux des bancs de filtres gamma (GFCCs) pour la reconnaissance de la parole de chuchotement en mode dépendant du locuteur. La plupart de ces études en plus que d'autres études tels que [Russo et al., 2019] démontrent que l'utilisation des caractéristiques à base des filtres gammatones donne une amélioration significative des performances dans des conditions bruitées.

Pour la reconnaissance du locuteur il existe aussi plusieurs travaux qui utilisent les filtres gammatones tel que le travail de [Al-Karawi, 2019] qui utilise une combinaison des caractéristiques MFCC et GFCCs pour la vérification du locuteur dans des conditions bruitées. De même pour Le travail de [Taherian et al., 2020] qui utilise les GFCC pour la reconnaissance du locuteur**.**

Approches basées sur les techniques de traitement d'image **:** L'idée de cette approche est de transformer le problème de la reconnaissance de l'audio en un problème de reconnaissance des images via la transformation du signal audio en une image à deux dimensions vu que le domaine de la reconnaissance des images est mieux étudié [Gouda et al., 2018]. Plusieurs travaux ont été réalisés et nous citons par exemple, le travail de [Gouda et al., 2018] qui propose un système de reconnaissance des commandes vocales en utilisant les CNNs (Convolutional Deep Neural Network) et en convertissant le signal audio en une image à deux dimensions. Un autre travail est celui de [Dennis et al., 2013a] qui propose l'utilisation des caractéristiques du spectrogramme pour la reconnaissance des événements acoustiques. D'autres travaux pour le même auteur peuvent être trouvés sur [Dennis et al., 2013b] et [Dennis et al., 2011]. Parmi aussi les travaux récents utilisant les techniques de reconnaissance des images pour la reconnaissance des sons, nous citons le travail de [Boddapati et al., 2017] qui propose un système de classification des sons de l'environnement où chaque événement audio est représenté sous forme d'images visuelles en les convertissant en Spectrogramme, MFCC et Cross Recurrence Plot (CRP). Les deux réseaux de neurones convolutionnels AlexNet et GoogLeNet sont utilisés.

- **Discussion**

La comparaison des travaux sur la RSE qu'on peut trouver dans la littérature est une tâche difficile vu que chaque travail s'intéresse à un type particulier de son (sons de pas, cris, ...), voir plusieurs sons mais qui différent d'un travail à un autre. L'auteur dans [Sehili, 2013] a essayé de mettre les différents travaux réalisés dans la reconnaissance de sons dans des catégories afin de montrer les vrais domaines par lesquels la RSE est influencée, autrement dit, la RSE utilise quelles méthodes ou approches ?

4 approches ont été citées dans sa synthèse : la reconnaissance de la parole, la reconnaissance du locuteur, le système auditif humain et les techniques de traitement d'image. Cependant, les quatre catégories ne sont pas indépendantes et elles peuvent être fortement liées. Prenons l'exemple des techniques de traitement d'image, cette nouvelle approche est très adoptée par la reconnaissance de la parole et du locuteur en même temps, et même chose pour les approches issues du système auditif humain. Par conséquent, il n'y a pas de variantes fixes pour pouvoir séparer et classer les approches qui peuvent être adoptées par la RSE. Pour cette raison, nous voyons donc important de se baser sur les méthodes de classification et d'extraction de caractéristiques utilisées par la RSE en mettant l'accent sur les taux de classification, les classes de sons visées ainsi que le nombre de classes pour pouvoir approximativement répondre à la question : quelles méthodes pour la RSE ?

2.3.2. Synthèse des travaux sur la RSE

Dans l'article [Chachada et Kuo, 2013] une étude approfondie sur le choix des caractéristiques pour la classification et la reconnaissance des sons environnementaux, cette étude a été basée sur la synthèse d'une bonne partie des travaux sur la RSE. La plupart de ces travaux utilisent l'MFCC soit seul car il donne les meilleures performances ou bien combiné avec d'autres caractéristiques pour améliorer la performance du système. De plus, un aspect notable dans cette étude est qu'il n'existe pas encore une façon claire pour le choix des caractéristiques pertinentes pour une application de RSE, ceci est dû premièrement à la non existence d'une base de données standard pour l'évaluation des solutions proposées, deuxièmement, le compromis entre la simplicité de la méthode d'un point de vue temps de calcul et l'efficacité de cette dernière. En générale, les méthodes stationnaires se caractérisent par leur simplicité et les méthodes non-stationnaires sont plus complexes mais plus performantes. Par conséquent, il faut classer les objectifs de l'application par ordre de priorité afin de pouvoir faire un bon choix des caractéristiques (système temps réel, application de surveillance, …).

De nombreux travaux de recherche ont été proposés dans le domaine de la reconnaissance de sons afin de reconnaître différentes catégories de sons qui dépendent de l'application visée. Dans cette partie, suite à la synthèse de [Chachada et Kuo, 2013], nous présentons une synthèse des travaux ainsi qu'une comparaison entre les travaux réalisés dans le domaine de la reconnaissance automatique de son. Notre comparaison est basée sur les méthodes d'extraction de caractéristiques utilisées ainsi que les méthodes de classification. Notre objectif étant d'identifier les méthodes les plus couramment utilisées et celles qui offrent les meilleurs taux de reconnaissance. Les méthodes de classification tournent en général autour des classifieurs GMM HMM et SVM vu qu'une autre section est dédiée pour une synthèse des travaux qui se basent sur les méthodes d'apprentissage profond. Dans ce qui suit, nous présentons une liste non exhaustive des travaux sur la RSE. En effet, cette synthèse est présentée en détail dans notre travail, qui se trouve dans [Abdoune et Fezari, 2017] :

Le travail présenté dans [Istrate, 2003] se concentre sur la recherche des paramètres acoustiques les plus efficaces pour un système de détection et de reconnaissance des sons pour la surveillance médicale. Les paramètres acoustiques classiques (*énergie du signal, LPC, LPCC, MFCC, LFCC, les dérivées des coefficients (*$\Delta, \Delta\Delta$)) sont d'abord testés, puis de nouveaux paramètres issus de la transformée en ondelettes ont été proposés. Le classifieur utilisé est un GMM. Le taux moyen d'erreur sur les sons purs est de 10% et sur les sons bruités à un RSB de 10 dB est de 22% et de 10% pour un RSB supérieur à 20 dB.

Le travail présenté dans [Chu et al., 2009] traite la reconnaissance des sons environnementaux pour la compréhension d'une scène ou le contexte entourant un capteur audio. La méthode MP (Matching Poursuit) a été choisie pour obtenir les caractéristiques du domaine fréquentiel-temporel les plus efficaces, car l'utilisation des caractéristiques du domaine fréquentiel uniquement échouent pour certains types de sons et plus particulièrement les sons ressemblant au bruit (ex. Sons de pluie, sons des insectes) disposant d'un large spectre plat. Pour la classification le GMM a été utilisé. Afin de montrer l'utilité des caractéristiques-MP des tests ont été faits sur les MFCC puis les caractéristiques MP et finalement la combinaison des MFCC et caractéristiques MP. Les résultats de l'application sont respectivement 75.3%, 84.0%, and 89.7%. Les résultats expérimentaux montrent des performances prometteuses dans le classement de 14 environnements audio différents.

Le même groupe dans un travail antérieur [Chu et al., 2006] dont la différence est la non utilisation des caractéristiques MP, ont trouvé les résultats suivants pour trois classifieurs différentes : 96.6% pour l'SVM, 94.3% pour le KNN, et 93.4% pour le GMM, en utilisant la sélection en avant des caractéristiques (forward selection of features). Les caractéristiques utilisées sont au nombre de 34 : 1^{er} – $12^{ème}$ MFCC, Écart-type des 1er – 12ème MFCC, Centroid spectral (Sc), Largeur de bande spectrale (Sw), Asymétrie spectrale (Sa), Planéité spectrale (Sf), ZC, Écart-type du ZC, Plage d'énergie (Er), Écart-type de la Plage d'énergie (Er), Fréquence de coupure (Roll-off), Écart-type du roll-off.

Dans [Muhammad et Alghathbar, 2009], les auteurs proposent une autre méthode pour la reconnaissance des environnements à partir de l'audio en combinant les MFCC, les descripteurs MPEG-7 et ZCR. L'utilisation complète des descripteurs MPEG-7 a montré une amélioration des performances par rapport à l'utilisation des MFCC. Le classifieur utilisé est l'HMM. L'expérimentation a montré que la combinaison de ces deux types de caractéristiques donne de meilleures performances par rapport à l'utilisation des MFCC seuls ou les descripteurs MPEG-7. Lorsque le ZCR est combiné avec ces deux derniers, une amélioration des performances a été observée pour certains types d'environnements. Ensuite, dans un autre travail [Muhammad et al., 2010], les mêmes auteurs proposent un système pour la reconnaissance des environnements en utilisant les descripteurs audio de bas niveau MPEG-7 avec les MFCC. La méthode FDR (Fisher Descriminant Rtaio) a été utilisée pour éliminer les descripteurs MPEG-7 non pertinents, ensuite l'ACP a été appliquée sur les 30 descripteurs

obtenus pour enfin arriver à 13 paramètres, ces derniers sont combinés avec les MFCC. Le classifieur utilisé est le GMM. Le système est évalué sur dix sons environnementaux différents. Le système proposé offre un taux de reconnaissance supérieur à celui des systèmes utilisant uniquement les MFCC ou les descripteurs MPEG-7 dans certains environnements. En somme, bien que les caractéristiques MPEG-7 surpassent les MFCC, leur combinaison améliore encore davantage le taux de reconnaissance.

Le même travail présenté dans [Muhammad et Alghathbar, 2009] et [Muhammad et al., 2010] a été présenté dans [AlQahtani et al., 2010] dont l'objectif est la reconnaissance des sons environnementaux en utilisant quelques descripteurs MPEG-7 et le ZC temporel, le classifieur utilisé est le KNN. Les performances sont évaluées par variation du nombre de fichiers d'apprentissage et le nombre d'échantillons par fichier. Les résultats ont montré que le bon taux de reconnaissance est obtenu par augmentation du nombre de fichiers d'apprentissage et en diminuant le nombre d'échantillons par fichier.

Ruben Delgado-Contreras et ses co-auteurs [Delgado-Contreras et al., 2014a] proposent une approche pour la classification des endroits par l'utilisation des 'empreintes digitales audio'. Les caractéristiques sont au nombre de 62 qui sont du domaine temporel, fréquentiel et statistique. Deux types de classifieurs ont été utilisés pour tester l'approche proposée : Random Forest et SVM. La base de son utilisée est collectée à partir d'une base de données collaborative en ligne appelée *Freesound*[2]. Le nombre de classes est 14 (14 environnements différents). Les résultats ont montré que le taux de classification est de 84.28% pour Random Forest et 91.42% pour les SVM.

Dans une autre expérimentation [Delgado-Contreras et al., 2014b], les auteurs utilisent une méthode de sélection des caractéristiques qui est '*Chi squared filter*' pour la classification des endroits. Les caractéristiques sont alors réduites de 62 à 15 (11 statistiques et 4 fréquentielles). Le classifieur utilisé est l'SVM, le nombre de classes est de 10 et le taux de reconnaissance est supérieur à 90%.

Dans le travail présenté dans [You et Li, 2012], une méthode appelée TESPAR (Time Encoded Signal Processing and Recognition) est proposée pour la reconnaissance des sons environnementaux. Cette méthode se caractérise par ses besoins en calcul qui sont petits par rapport aux autres méthodes, et elle a été testée sur une base de données extraite à partir de la base Freesound. Afin d'évaluer le système proposé, une comparaison a été faite avec un système à base d'MFCC et un classifieur SVM sur la même base de données. Les résultats ont montré que TESPAR est plus efficace dans l'existence du bruit et son temps de calcul est trop petit par rapport à l'SVM. Les taux de classification obtenus dans un environnement non bruité sont 94% pour TESPAR et 98% pour l'SVM et ceci pour le premier test, et pour plus de détails sur les différents tests effectués se référer à l'article.

[2] http://www.freesound.org

Le travail présenté dans [Dufaux et al., 2000] traite la détection et la reconnaissance automatique des sons impulsifs comme les bris de verre, les cris, l'explosion. Le système a été évalué sur une base de données qui contient 822 signaux composant 6 classes différentes. L'algorithme de détection est basé sur les *filtres médians* analysant les variations d'énergie et il offre de bonnes performances même dans le bruit. Deux classifieurs statistiques ont été utilisé pour la classification le GMM et l'HMM afin de comparer les résultats. Les résultats ont montré que le taux de reconnaissance est de 98% pour un SNR de 70dB et il est moins de 80% pour un SNR de 0%.

Vacher et ses co-auteurs dans [Vacher et al., 2010a], proposent un système de reconnaissance sonore complet appelé AuditHIS pour identifier les différents sons présents dans l'appartement afin de reconnaître les activités effectuées de la vie quotidienne. Ce dernier est associé à un système de reconnaissance vocale en français appelé RAPHAEL pour rechercher des mots-clés de détresse à l'intérieur du signal mesuré. Le traitement et l'analyse du son passent par les étapes suivantes : après *Acquisition et prétraitement* vient l'étape de *Détection* qui est l'estimation du début et la fin du son de chaque événement audio. Ensuite, dans la phase de *Segmentation* un classifieur GMM est utilisé *(classifieur GMM utilisant **24 modèles gaussiens*** et dont les caractéristiques acoustiques sont les ***LFCC avec 16 filtres**)* où chaque signal audio est classifié comme parole ou son de la vie courante. Finalement, vient l'étape de *Classification du son* ou reconnaissance de la parole qui détermine quelle classe du son ou quelle phrase a été prononcée. Chaque son de la vie courante est classifié soit par un classifieur GMM (bons résultats lorsque le SNR (Signal to Noise Ratio) est moins de 10 dB) ou HMM (meilleurs résultats dans un environnement non bruité). L'apprentissage a été effectué sur un corpus qui contient 8 classes. Les performances globales du système sont 89,76% de sa bonne différenciation son/ parole et 72,14% des sons bien-classés.

Le travail de [Rabaoui et al., 2007], traite la classification supervisée des signaux sonores pour une application de télésurveillance. La méthode de classification utilisée est plusieurs SVM à une classe (1-SVM). Les descripteurs audio utilisés sont : Discrete Wavelet Transform (DWT), MFCC, Energie, Log énergie, Spectral Roll-off point (SRF), Spectral Centroid (SC) et le ZCR. Le taux de la bonne classification est 96.89%. Les sons à reconnaitres sont : Cris de secours, Coups de fusils, bris de verre, explosions, claquements de portes, Aboiement de chiens, sonneries de téléphones, voix d'enfants, et sons de machines.

Dans ce travail [Uzkent et al., 2012], les auteurs proposent un système de classification des sons environnementaux non-parole en utilisant un nouvel sous ensemble de caractéristiques 2D, utilisé avec une méthode d'extraction des caractéristiques basée sur le pitch (pitch range (PR)) et appelé descripteurs PR. Trois classifieurs sont utilisés pour évaluer les performances du système, un SVM (SVM avec noyau linéaire et SVM avec un noyau gaussien), un réseau de neurones à fonctions de base radiales (RBF) et un classifieur basé sur la méthode des plus proches voisins. En ce qui concerne les SVM, les trois méthodes OAO, OAR, and DDAG ont été utilisées afin de transformer l'SVM binaire à un SVM multi-classes et ceci avec l'outil LIBSVM,

avec un noyau linéaire et gaussien pour le classifieur SVM. Les sons à reconnaitre sont : Les coups de feu, bris de verre, les cris, les aboiements de chiens, la pluie, les sons des moteurs, et le bruit de restaurant. Une comparaison a été faite avec le même système mais en utilisant les MFCC comme descripteurs. Les résultats de cette expérimentation montrent que les meilleurs taux de bonne reconnaissance, pour les trois types de classifieurs, sont obtenus par la combinaison des descripteurs PR et MFCC, et les résultats obtenus par les MFCC sont mieux que ceux obtenus par les PR. En ce qui concerne les classifieurs, les SVM avec un noyau linéaire donnent un taux moyen de reconnaissance de 85,6% pour la combinaison (PR et MFCC), les SVM à un noyau gaussien donnent un taux de 88,7%, les réseaux de neurones RBF 81,78%, et finalement le classifieur NN donne un taux de 86,4%. Par conséquent, dans cette expérimentation nous constatons que l'SVM avec un noyau gaussien est le meilleur classifieur et que les descripteurs PR combinés avec les MFCC donnent les meilleurs résultats.

Dans [Hang et al., 2019], trois méthodes de classification sont sélectionnées pour la reconnaissance des sons d'eau, à savoir SVM, KNN et CNN. Les caractéristiques utilisées sont extraites des empreintes audio (20 caractéristiques). Les résultats expérimentaux ont montré que le taux de reconnaissance obtenu par les SVM qui est 98.22% est supérieur aux taux obtenus par les deux autres modèles de classification qui sont 97.75% et 70.29% pour le KNN et le CNN respectivement.

Enfin, un travail récent dans [Jesudhas et Ranjan, 2024] utilise l'SVM pour la classification des sources sonores dans un environnement domestique, au lieu des solutions basées sur des modèles d'apprentissage profond, afin d'obtenir une haute précision avec des coûts de complexité réduits. Des caractéristiques telles que la dispersion spectrale et les GTCC (Coefficients Cepstraux de Gammatones) sont extraites, avec une précision de 80% en phase de validation et de 60% en environnement temps réel.

Le tableau 2.1 présenté ci-dessous résume les travaux montrés précédemment, en mettant l'accent sur les descripteurs utilisés, les méthodes de classification et les résultats obtenus.

Tableau 2. 1. Synthèse des travaux sur la RSE en précisant les méthodes d'extraction de caractéristiques utilisées et les méthodes de classification

Travail (auteurs/référence)	Objectif	Caractéristiques utilisées	Nombre de caractéristiques	Méthode de sélection des paramètres	Classifieur	Précision
[Chu et al. , 2009]	la compréhension d'une scène	– Les caractéristiques MP – Les caractéristiques MP + MFCC – MP MFCC (16), MFCC (12), ΔMFCC (12), LPC (12), ΔLPC (12), LPCC(12), …	14	MP (Mutching Pursuit)	GMM	-MP : 75.3%, -MFCC : 84.0%, -MP+ MFCC : 89.7%
[Chu et al.,	Reconnaissance des	1er – 12ème MFCC, Écart-type des 1er – 12ème MFCC, Centroid	34	Sélection	SVM	96.6%

2006]	environnements pour les robots mobiles	spectral (Sc), Largeur de bande spectrale (Sw), Asymétrie spectrale (Sa), Planéité spectrale (Sf), ZC, Écart-type du ZC, Plage d'énergie (Er), Écart-type de la Plage d'énergie (Er), Fréquence de coupure (Roll-off), Écart-type du roll-off.		en avant	KNN GMM	94.3% 93.4%
[Ruben Delgado-Contreras et al. , 2014b]	Classification des endroits	- *Caractéristiques temporelles* : Taux de passage par zéro à court terme (Short-Time Average Zero-Crossing Rate), Énergie logarithmique à court terme, Énergie quadratique à court terme, etc. *-Caractéristiques fréquentielles* : Flux Spectral, Spectral Roll-Off, Centroid Spectral, ...) *-Caractéristiques statistiques* :1er et 2ème ordre : Maximum, Minimum, Moyenne, Médian, Déviation standard, Variance, ...	62	non	Random Forest SVM	84.28% 91.42%
[Ruben Delgado-Contreras, 2014a]	Classification des endroits	11 statistiques et 4 fréquentielles	15	Chi squared filter	SVM	Supérieur à 90%
[Muhammad et Alghathbar , 2009]	Reconnaissance des environnements	MFCC, les descripteurs MPEG-7 et ZCR		PCA	HMM	Le taux moyen n'est pas calculé
[Muhammad et al. , 2010]	Reconnaissance des environnements	Decsripteurs MPEG-7 et MFCC	13 descripteurs MPEG	FDR ACP	GMM	Entre 90 et 96 % (dépends de la classe)
[Dufaux et al., 2000]	Détection et reconnaissance des sons impulsifs	L'énergie	1	non	GMM HMM	98% à 70dB au dessus de 80% for 0dB
[Vacher et al., 2010a]	Reconnaissance des sons de la vie courante (8 classes)	Les paramètres LFCC (16 filtres)	--	non	GMM (24 modèles gaussiens)	72.14%
[Rabaoui et al. , 2007]	Classification des sons de	DWC + MFCC + Energy + Log energy + SRF + SC + ZCR	7	non	1-SVM	96.89%

	l'environnement					
[Uzkent et al., 2012]	Classification des sons de l'environnement	Descripteurs PR(à base de pitch) + MFCC		non	-SVM à noyau gaussien, -ANN (RBF) -KNN.	-88,7% (SVM avec noyau gaussien) -85,6% (SVM avec noyau linéaire) -81,78%, RN (RBF) -86,4%. KNN
[Hang et al., 2019]	reconnaissance des sons d'eau	20 caractéristiques extraites des empreintes audio	20	non	SVM KNN CNN	-98.22% (SVM à noyau RBF) -97.75% (KNN) -70.29% (CNN)
[Jesudhas et Ranjan, 2024]	Classification des sons environnementaux	Les coefficients GTCC, spectral spread, spectral roll-off, spectral flux, ..	--	non	SVM	80% dans la validation 60% dans un environnement temps réel

- **Discussion**

Deux points importants qu'on peut discuter dans cette étude : les méthodes d'extraction des caractéristiques et les méthodes de classification adoptées dans chacun de ces travaux. Commençant tout d'abord par les méthodes d'extraction de caractéristiques.

A partir de cette synthèse nous concluons que la plupart des travaux de détection ou de reconnaissance des sons de l'environnement cités dans la littérature utilisent les MFCC. Si ce dernier n'est pas utilisé seul il est combiné avec d'autres paramètres pour améliorer la performance du système. Les caractéristiques du domaine temporel-fréquentiel sont aussi très utilisés pour la reconnaissance des sons environnementaux. D'autres paramètres aussi tels que les descripteurs MPEG-7 sont utilisés et ils deviennent plus efficaces lorsqu'ils sont combinés avec les MFCC. De plus, d'autres études comparatives comme dans [Chu et al., 2009] et [Sharan et Moir, 2016], affirment que les MFCC fonctionnent bien pour les sons structurés tels que la parole et la musique [Chu et al., 2009], ils sont aussi les plus utilisés dans les applications de reconnaissance de la parole et de son [Sharan et Moir, 2016], mais leur performance se dégrade en présence du bruit [Chu et al., 2009] [Sharan et Moir, 2016]. Les MFCC ne sont pas efficaces pour l'analyse des signaux de type bruit ayant un spectre plat. L'environnement audio contient une grande et diverse variété de sons, comme pépiements des insectes et des sons de pluie qui sont généralement de type bruit avec un large spectre plat qui ne peuvent pas être efficacement modélisés par les MFCC [Chu et al., 2009].

Concernant les méthodes de classification, nous remarquons aussi que cette étude a porté uniquement sur trois types de classifieurs : les GMM, les HMM et les SVM, dans le but de les comparer en terme des taux de reconnaissance, sans oublier quelques travaux qui ont testé aussi les KNN et ANN. Une motivation pour cette comparaison est le choix de la méthode appropriée à notre application et à nos besoins. A partir de cette synthèse nous constatons que la plupart des travaux utilisent l'SVM et il s'avère la méthode la plus puissante lorsqu'il est comparé avec d'autres méthodes classiques. L'SVM avec un noyau gaussien montre aussi un résultat très satisfaisant en comparaison avec d'autres noyaux tel que le noyau linéaire. Enfin, les méthodes de sélection des paramètres acoustiques ne sont pas utilisées dans la plupart des travaux cités, sauf pour les travaux utilisant un grand nombre de caractéristiques.

En effet, Non seulement l'utilisation des SVM par la plupart des travaux de RSE qui nous a incité à utiliser cette méthode dans notre thèse, mais aussi son fondement théorique solide et sa capacité de discrimination et de généralisation tel qu'expliqué dans le chapitre précédant sont une cause aussi pour choisir l'SVM comme première solution pour notre système de classification des sons de la vie courante.

Une question importante qui nous vient à l'esprit : Que disons-nous sur les nouvelles méthodes de classification de deep learning tels que les réseaux de neurones convolutionnels ? Sont-elles appliquées pour la RSE ? La section suivante, par conséquent, est consacrée pour la présentation et la discussion des travaux basés sur ces nouvelles méthodes.

2.4. Travaux de RSE à base du deep learning

Au début de notre travail, la recherche d'une méthode d'apprentissage automatique classique tel que GMM, SVM, HMM pour la RSE était notre première préoccupation, où nous nous sommes attirés par les SVM pour plusieurs raisons tels que leur capacité de généralisation même avec un ensemble d'apprentissage limité, leur capacité aussi à définir des frontières de séparation entre les classes de façon optimale, en plus de leur taux d'utilisation dans les domaines de la reconnaissance de sons qui sont cités dans les sections précédentes. Par la suite, avec l'émergence des méthodes d'apprentissage profond connus sous le nom 'Deep Learning', vu l'augmentation des puissances de calcul et le développement d'algorithmes puissants, il s'avère important de faire partie de cette nouvelle méthode dans ce manuscrit en montrant d'abord son utilisation dans le domaine de la reconnaissance des sons, les performances atteintes, puis la comparer avec les premières méthodes.

En réalité, plus récemment, le Deep Learning, c'est-à-dire les réseaux de neurones artificiels avec plus d'une couche intermédiaire cachée, a gagné en popularité et a obtenu des résultats impressionnants sur plusieurs tâches d'apprentissage automatique [LeCun et al., 2015], [Sigtia et al., 2016]. Il a remporté un grand succès dans de nombreux domaines tels que le traitement du langage naturel, la reconnaissance vocale, la vision par ordinateur, l'analyse d'images et vidéos et le multimédia. Aujourd'hui, plusieurs travaux utilisent le deep learning

pour la reconnaissance automatique des sons. De ce fait, nous présentons dans ce qui suit, quelques travaux récents sur la RSE qui se basent sur le deep learning (voir tableau 2.2).

Sigtia dans [Sigtia et al., 2016] propose un système de détection des cris de bébés et des alarmes de la fumée en utilisant les réseaux de neurones profonds (DNN) puis compare les résultats avec les GMM et les SVM. Les DNN offrent un meilleur taux de précision puis les SVM et ensuite les GMM.

Un autre travail dans [McLoughlin et al., 2017] traite la détection et la reconnaissance d'un flux audio continu dans des conditions bruitées en utilisant les méthodes du deep learning. Pour des besoins de comparaison, les classifieurs suivants ont été testés sur la base de données : HMM avec MFCC, SIF (spectrogram image feature) avec SVM, SIF avec DNN, SIF avec CNN. Dans le cas des sons isolés, le système SIF-CNN fonctionne le mieux, suivi par SIF-SVM puis SIF-DNN. Le HMM est le moins robuste au bruit. Dans le cas d'un flux audio continu, les performances SVM sont très compétitives par rapport au système CNN dans tous les cas, plus que le DNN en fait.

Nanni et ses co-auteurs dans [Nanni et al., 2017] proposent un système de détection pour aider à l'auscultation des sons cardiaques. Deux modèles ont été testés dans cette étude notamment, le CNN et le CNN combiné avec LSTM. Les taux de précision obtenus pour les deux modèles sont 93,07% et 91,06% respectivement.

L'étude présentée dans [Kim et al., 2020] traite les événements d'urgence qui ont un impact critique sur la santé des occupants et propose un modèle de reconnaissance sonore basé sur l'apprentissage en profondeur pour surveiller les comportements des occupants et détecter d'éventuels événements d'urgence. Deux modèles de classification ont été utilisés : CNN et LSTM (Long Short-Term Memory). En effet, le LSTM développé dans cette recherche est un modèle RNN (réseaux de neurones récurrents) plus avancé qui a pour objet de palier aux défauts des RNN traditionnels. Les expériences sont menées à l'aide de données audio collectées à partir d'environnements domestiques SPH (Single-Person Households) réels et de sites Web de partage de données en ligne. Les résultats expérimentaux ont démontré que le modèle développé pouvait distinguer avec succès les événements sonores d'urgence des sons des activités humaines régulières. Les résultats expérimentaux ont démontré que le CNN surpassait le modèle LSTM dans la classification des événements sonores d'urgence ainsi que dans la surveillance du comportement des occupants, avec des taux de précision de 83.9% pour le CNN et de 62.6% pour le LSTM.

Dans [Greco et al., 2020] les auteurs proposent une méthode d'apprentissage en profondeur baptisée AReN (Audio Event Recognition Network) à 21 couches pour reconnaître automatiquement les événements d'intérêt dans le contexte de la surveillance audio à savoir les cris, les bris de verre et les coups de feu. Les signaux en entrée sont représentés par une image gammatonegram ; il s'agit d'un spectrogramme basé sur une bande de filtre gammatone. Le système a été testé sur trois bases de données différentes à savoir SESA,

MIVIA des événements audio et MIVIA des événements routiers, les taux de reconnaissance atteints sont 91.43%, 99.62% et 100% respectivement. Une comparaison été faite dans cette étude entre les méthodologies traditionnelles d'apprentissage automatique et l'apprentissage profond et cette comparaison confirme l'efficacité de l'approche proposée.

Enfin, un travail aussi plus récent est celui de [Lee et al., 2021] qui vise la réalisation de deux sous tâches : la classification de l'audio en 10 classes et la deuxième consiste à classifier l'audio en trois catégories en se basant sur des solutions à faible complexité. Dans ce travail, les paramètres Deltas-DeltaDeltas et HPSS ont été utilisés avec quatre modèles de classifieur inspirés du VGGNet, ResNet, LCNN, et InceptionNet. Dans les quatre modèles, l'utilisation de Deltas-DeltaDeltas a surpassé les performances de HPSS, et parmi eux, l'utilisation de ResNet a présenté la précision la plus élevée. Dans la sous-tâche B, l'utilisation de ResNet réduit en utilisant la fonction Deltas-DeltaDeltas a présenté les meilleures performances à 95,38 %.

Tableau 2. 2. Systèmes de reconnaissance de sons basés sur les méthodes du deep learning

travail	objectif	Méthode de classification	Description des résultats
[Sigtia et al., 2016]	Détection des cris de bébés et des alarmes de la fumée	-DNN -GMM -SVM	Les DNN offrent le meilleur taux de reconnaissance puis l'SVM puis le GMM
[McLoughlin et al., 2017]	Détection et reconnaissance d'un flux audio continu dans des conditions bruitées	-HMM avec MFCC -SIF avec SVM -SIF avec DNN -SIF avec CNN	-performances SVM sont très compétitives par rapport au système CNN plus que le DNN - Le HMM est le moins robuste au bruit
[Nanni et al., 2017]	Détection des sons pour aider à l'auscultation des sons cardiaques	-CNN -CNN combiné avec LSTM (Long Short-Term Memory)	93,07% 91,06%
[Kim et al., 2020]	Détection des évènements d'urgence	CNN LSTM	83.9% pour le CNN et de 62.6% pour le LSTM
[Greco et al., 2020],	La surveillance audio	AReN (Audio Event Recognition Network)	91.43%, 99.62% et 100% sur trois base de données différentes

[Lee et al., 2021]	-Classification des sons en 10 classes -Classification des sons en trois catégories	Les paramètres Deltas-DeltaDeltas et HPSS avec : VGGNet, ResNet, LCNN, et InceptionNet	ResNet et Deltas-DeltaDeltas a présenté les meilleures performances à 95,38 %.

- **Discussion**

Dans cette partie, nous avons présenté un aperçu sur quelques travaux de détection et reconnaissance de sons en utilisant les méthodes du deep Learning. Malgré qu'un nombre restreint des travaux a été présenté dans cette section, mais nous constatons l'importance de cette méthode et sa puissance lorsqu'elle est comparée avec les méthodes traditionnelles tels que les HMM, GMM et SVM. Dans la plupart des travaux, le taux de reconnaissance obtenu par ces méthodes du deep learning est le plus élevé. Sachant qu'il existe différentes méthodes du deep learning, mais chaque méthode donne un résultat différent qui peut être meilleur par rapport aux résultats obtenus par les autres méthodes classiques ou bien l'inverse. Par conséquent, nous voyons l'intérêt d'introduire ces méthodes dans nos futurs tests et travaux.

2.5. Travaux sur la détection des situations de détresse

Il existe plusieurs travaux sur la télésurveillance audio. Dans ce paragraphe nous présentons quelques travaux ou systèmes de télésurveillance en précisant surtout la nature de l'environnement, le nombre de classes et les types de sons étudiés ainsi que la ou les méthodes de classification utilisées et éventuellement le taux de bonne reconnaissance.

Commençons par le système de télésurveillance proposé dans [Radhakrishnan et al., 2005]. Ce travail présente une solution hybride pour la télésurveillance dont l'objectif est la détection des délits dans un ascenseur. Le système proposé est composé de deux sous-systèmes : le premier est un classifieur supervisé et le deuxième est un analyseur audio non supervisé. Le but de ce dernier est la détection d'autres sons suspects non pris en charge par le premier sous-système avec possibilité de mettre à jour le model en ajoutant de nouvelles classes de sons suspects, ce système est basé sur le GMM. La base de données est composée de *quatre classes* : claquement, sons de pas, discours non neutre et discours normal. Les paramètres utilisés sont 12 MFCC pour un frame de 8 millisecondes, l'étude a été faite sur 126 clips audio avec des sons suspects et 4 clips sans événements. Le taux de reconnaissance obtenu par le GMM est de 85%.

Un autre travail [Rouas et al., 2006], décrit aussi un système de télésurveillance audio dans un véhicule de transport en commun qui est un environnement bruité. Différents sons peuvent se produire, les auteurs sont basés sur les 5 scénarios ou sons suivants : bagarre entre deux hommes ou plus, entre deux femmes ou plus, entre hommes et femmes, vol à main armée, et vol de sac. L'expérimentation se limitait à la détection des cris. Les paramètres

acoustiques utilisés sont les MFCC, LPC, l'énergie, et PLPC, et pour la classification l'SVM et le GMM ont été utilisés pour comparer les performances et détecter les cris dans cet environnement. Le taux de reconnaissance atteint est 75% pour la détection des cris, et 98% pour la détection des événements qui ne contiennent pas les cris. L'expérimentation a montré que le classifieur SVM avec l'utilisation des PLP a donné les meilleures performances.

Un autre travail est celui de [Valenzise et al., 2007], qui décrit un système de surveillance audio pour la détection et la localisation des événements anormaux dans un endroit public tels que les cris, les coups de feu. Ce système utilise deux classifieurs GMM qui travaillent en parallèle pour la discrimination respectivement, des cris du bruit et les coups de feu aussi du bruit. Le nombre de caractéristiques utilisées est de 13 pour la classification cris/bruit et 14 pour la classification coups de feu/bruit et ceci après une étape de sélection des paramètres. Une précision de 93% a été obtenue avec un taux de faux rejet de 5% lorsque le RSB est de 10 dB.

Dans une autre tendance [Ntalampiras et al., 2008], un autre système de détection des situations de détresse dans un endroit public a été présenté. Les classes de sons au nombre de *trois* : cris, bruit, et coups de feu. Ce système utilise deux GMM binaires en parallèle qui permettent respectivement de distinguer entre cris et bruit et entre coup de feu et bruit. Chaque frame est classifié par les deux classifieurs en même temps. La décision finale est prise par calcul du OU logique. Les paramètres utilisés sont de différents types : temporels, spectrales, perceptuels et de corrélation.

Les auteurs dans [kuklyte et al., 2009], ont étudié les événements anormaux dans un environnement bruité en utilisant les MFCC et les HMM. *Quatre classes* ont été visées notamment l'explosion, les coups de feu, les cris comme sons anormaux ou de détresse et le bruit de métro comme événement normal. Le taux de bonne classification est 93.3%.

Dans [Foggia et al., 2016], les auteurs proposent un système de détection des accidents de la route par identification des situations dangereuses telles que le dérapage des pneus et les accidents de voiture. L'SVM avec un noyau linéaire a été utilisé et après une étude sur les paramètres acoustiques à utiliser les résultats ont montré que les paramètres MFCC et Bark sont les meilleurs pour différents SNR. La précision moyenne est de 78.95% à une distance maximale de 120 mètres.

Les auteurs dans [Cakir et al., 2017], proposent d'appliquer un CRNN (une combinaison de CNN et de RNN) à une tâche de détection des événements sonores (SED) polyphonique. Le terme polyphonique veut dire que le système peut traiter le cas de l'existence de plusieurs sons en même temps contrairement au mot monophonique. La métrique d'évaluation est le *score F1* ainsi que les vrais positifs, les faux positifs, et les faux négatifs. Les résultats fournis par le système sont prometteurs pour les quatre bases de données testées. De plus, Les résultats montrent une amélioration significative avec l'introduction de méthodes d'apprentissage en profondeur. CRNN présente des performances nettement supérieures à

celles des méthodes précédentes (HMM, GMM) et présente encore une amélioration considérable par rapport aux autres approches utilisant les réseaux de neurones.

Enfin, Dans un travail plus récent [Min et al., 2019], l'auteur présente un système de détection des événements d'urgence à l'intérieur en utilisant les CNN. Les sons en question sont : les sons indiquant des événements d'urgence (explosion, coup de feu, bris de verre et cri) et un seul son normal (le sommeil). Les paramètres acoustiques calculés sont les spectrogrammes mel à l'échelle logarithmique (log-scaled mel-spectrograms). L'expérimentation a donné comme résultat un F-score de 77.32%.

Le tableau 2.3 ci-dessous résume l'ensemble de ces travaux.

Tableau 2. 3. Systèmes de détection des situations de détresse

Référence	environnement	Les classes	Méthodes
[Radhakrishnan et al., 2005]	L'ascenseur	claquement, sons de pas, discours non neutre et discours normal	MFCC et GMM
[Rouas et al., 2006]	véhicule de transport en commun	Les cris	MFCC, LPC, l'énergie, et PLPC, et pour la classification l'SVM et le GMM
[Valenzise et al., 2007]	endroit public	cris, les coups de feu	GMM
[Ntalampiras et al., 2008]	endroit public	cris, bruit, et coups de feu.	GMM
[kuklyte et al., 2009]	métro	l'explosion, les coups de feu, les cris et bruit métro	les MFCC et les HMM
[Foggia et al., 2016]	La route	le dérapage des pneus et les accidents de voiture	MFCC et Bark et SVM
[Cakir et al., 2017]	Environnement extérieur et intérieur (maison)	Différentes BDD et différentes classes	CRNN
[Min et al., 2019]	(intérieur) La maison	explosion, coup de feu, bris de verre et cri	CNN

- **Discussion**

Dans cette section nous avons exploré quelques travaux de détection des situations de détresse via la reconnaissance des sons anormaux indiquant un danger tels que les coups de feu, que ce soit à l'intérieur ou à l'extérieur. Nous voyons l'intérêt de ces applications dans la

vie quotidienne qui est toujours la surveillance. La reconnaissance de quelques classes de sons est un aspect commun pour ces différents travaux. Différentes bases données ont été utilisées pour l'évaluation des systèmes proposés. Les méthodes de classification utilisées sont diverses, nous citons les GMM, HMM, SVM, et les méthodes du deep learning. Les taux de reconnaissance obtenus varient d'une application à une un autre et d'une méthode à l'autre dans le même travail.

L'objectif de cette section est d'examiner et de comparer les méthodes utilisées pour la reconnaissance des sons environnementaux et celles utilisées pour la reconnaissance des situations de détresse et de danger. En effet, nous avons démarré du principe est ce que des sons particuliers utilisent des méthodes particulières ? en d'autres termes, est ce que cette catégorie de sons anormaux comme les coups de feu et les cris sont bien reconnues en utilisant quels paramètres acoustiques et quelles méthodes de classification ?

Nous constatons qu'il n'existe pas de méthodes spéciales pour l'extraction des caractéristiques et la classification des sons de détresse, et que les mêmes méthodes que celles utilisées pour la reconnaissance des sons environnementaux peuvent être utilisées pour la détection de la détresse.

2.6. Conclusion

Dans ce chapitre, nous avons d'abord présenté et introduit le domaine de la reconnaissance des sons et les différents concepts liés à ce domaine notamment l'analyse de scène auditive (ASA) ensuite, nous avons parlé des différentes situations qui peuvent avoir lieu dans l'ASA tels que le cocktail party qui est la survenue de plusieurs sons en chevauchement en même temps. En effet, la reconnaissance de sons forme une étape dans le processus entier de l'analyse de scène auditive, cette étape tient place toujours après séparation des sons en entrée. Par la suite, nous avons abordé différents travaux sur la reconnaissance automatique de sons qui varient par leurs méthodes et les bases de données utilisées pour l'évaluation des performances.

Dans une **première étude** comparative des *systèmes de reconnaissance* de sons, nous avons mis l'accent sur les méthodes de classification utilisées, les méthodes d'extraction de caractéristiques, les bases de données utilisées et les taux de reconnaissance. Cette comparaison a touché des applications dont les objectifs tournent en général autour de la RSE pour la reconnaissance des endroits, et la reconnaissance des sons de la vie courante dans des environnements indoor. A partir de cette synthèse, nous pouvons tirer les conclusions suivantes :

- La majorité des systèmes de reconnaissance de son utilisent l'MFCC ou les combinent avec d'autres paramètres et montrent toujours une amélioration des performances.
- Les MFCC sont des caractéristiques du domaine temps-fréquence et elles sont particulièrement précieuses dans la reconnaissance des sons environnementaux car

elles capturent à la fois des informations dans les domaines temporel et fréquentiel, permettant une analyse plus complète des modèles sonores dans le temps et sur différentes gammes de fréquences. Cependant, Les MFCC ne sont pas efficaces pour l'analyse des signaux de type bruit ayant un spectre plat.

- Les méthodes de classifications communément utilisées pour la RSE sont en général, les HMM, les GMM et les SVM.
- La plupart des travaux utilisent l'SVM et il s'avère la méthode la plus puissante lorsqu'il est comparé avec d'autres méthodes classiques.
- Nous synthétisons également, qu'une caractéristique importante des sons non-parole est leur diversité, et les systèmes existants sont spécialisés traitant un type particulier de son ou un ensemble de sons mais avec difficulté de traiter toutes les catégories de sons, ce qui a rendu difficile de trouver les bonnes méthodes pour la RSE, et a rendu aussi la comparaison des système une tâche complexe. En conséquent, ce domaine reste toujours ouvert pour trouver des solutions, comparer et essayer de lui adapter les méthodes existantes et même de trouver ses méthodes.

La **deuxième étude** a porté sur l'analyse et la comparaison des travaux de *détection des situations de détresse* par la reconnaissance des sons anormaux tels que les cris, les coups de feu. L'objectif principal de ces applications est la surveillance. Chaque travail définit ses sons d'intérêt, autrement dit, les sons indiquant une situation de danger. Différentes bases de données ont été utilisées pour chaque travail et les méthodes de classification aussi varient d'une application à une autre et il n'y a pas une méthode bien précise pour ce type d'application. Les taux de reconnaissance obtenus varient d'une application à une un autre et d'une méthode à l'autre dans le même travail.

En effet, si nous voulons juger l'efficacité des méthodes de classification pour la détection des sons anormaux ou de détresse, le taux de reconnaissance ou la précision constitue le seul moyen. Cependant, si on veut comparer un travail avec un autre ceci devient une tâche difficile, du fait que d'un côté, chaque travail utilise sa propre base de données ou bien une base de données différente, de l'autre côté, les classes de sons à reconnaitre varient d'une application à l'autre malgré que l'objectif étant toujours le même. Le problème posé par le choix des méthodes de classification est le compromis entre la précision et le coût de calcul. Des méthodes qui donnent de bons taux de reconnaissance ne présentent pas le bon choix lorsque leur temps d'exécution est important, le cas d'un système qui doit répondre en temps réel. Sigtia et ses co-auteurs dans [Sigtia et al., 2016], par exemple, comparent les performances de DNN avec les GMM et SVM sur une tâche de détection d'événement audio environnemental en tenant en compte l'aspect puissance ou coût de calcul. Les résultats ont montré que DNN offrent le meilleur rapport de précision de classification sonore sur une gamme de coûts de calcul, tandis que les GMM offrent une précision raisonnable à un coût constamment faible, et les SVM se situent entre les deux en termes de compromis entre précision et coût de calcul. De plus, Les systèmes RSE sont généralement déployés sur du

matériel embarqué, ce qui impose de nombreuses contraintes de calcul [Sigtia et al., 2016]et [Krstulović, 2018]. Le travail de [Jesudhas et Ranjan, 2024] publié récemment, aussi affirme que les modèles de classification en apprentissage automatique traditionnel sont mieux adaptés aux dispositifs en périphérie (edge devices) que les réseaux de neurones convolutifs (CNN).

La troisième étude est un aperçu plutôt qu'une étude des *travaux de reconnaissance de sons basés sur les méthodes du deep learning*. Selon les travaux présentés dans cette section, nous voyons l'importance de ces méthodes dans l'amélioration des taux de reconnaissance par rapport aux méthodes classiques. Cependant, il existe des cas où des méthodes classiques puissantes tels que les SVM donnent des résultats similaires et même plus performants. Sachant qu'il existe différentes méthodes du deep learning, mais chaque méthode donne un résultat différent qui peut être meilleur par rapport aux résultats obtenus par les autres méthodes classiques ou bien l'inverse. Nous constatons conséquemment, l'intérêt d'introduire ces méthodes dans nos futurs tests et travaux mais en tenant compte de quelques critères tels que leur nécessité de grandes bases de données pour l'apprentissage et leurs besoins intensifs en calcul.

En résumé, à partir des différentes synthèses que nous avons présentées dans ce chapitre nous constatons l'intérêt des méthodes de classification traditionnelles telles que SVM et GMM pour la reconnaissance du son. L'SVM en particulier, montre son large utilisation dans la RSE et il s'avère la méthode la plus puissante. De plus, les SVM sont cruciales pour la reconnaissance des sons environnementaux en raison de leur forte capacité à distinguer les différentes classes sonores, même dans des espaces de caractéristiques de grande dimension. Aussi, ils sont efficaces pour traiter des données complexes et non linéaires que l'on trouve souvent dans les données sur les sons environnementaux en raison de leur capacité à maximiser la marge entre les classes. De plus, les SVM sont robustes avec des ensembles de données plus petits, atteignant souvent une grande précision sans avoir besoin de données étendues, ce qui les rend idéales pour les applications pratiques de reconnaissance sonore.

D'un autre côté, les méthodes de classification basées sur l'apprentissage profond montrent également des résultats très satisfaisants, dépassant la plupart du temps ceux obtenus par les méthodes traditionnelles, mais le SVM reste également un concurrent pour ces classificateurs. En effet, Les systèmes AESR sont généralement déployés sur du matériel embarqué, ce qui impose de nombreuses contraintes de calcul [Krstulović, 2018]. Cependant les méthodes basées sur le deep learning nécessitent beaucoup de calculs et leur déploiement dans des applications en temps réel nécessite une puissance de traitement élevée. En outre, ils requièrent de grands ensembles de données pour l'apprentissage. Finalement, concernant les méthodes d'extraction de caractéristiques, les MFCC aussi constituent la base de la plupart des travaux de RSE vu leur capacité à capturer à la fois des informations dans les domaines temporel et fréquentiel, et donnent des résultats satisfaisants.

En conclusion, Vu la nature de notre application de départ où l'objectif est de doter la maison par un système de télésurveillance via l'analyse des sons produits dans l'appartement qui sont des sons de la vie quotidienne, et d'après les nos conclusions précédentes, le choix des SVM et des MFCC représente initialement la meilleure option compte tenu des critères et des objectifs de notre application et du fait que le problème posé par le choix des méthodes de classification est le compromis entre la précision et le coût de calcul.

CHAPITRE 3
Corpus de sons de la vie courante

Ce chapitre décrit le corpus de sons de la vie courante, les classes de sons, les conditions et paramètres d'enregistrement, etc. Pour cette raison des travaux antérieurs sur la création des bases de données ainsi que les travaux utilisant des bases de données existantes vont être présentés au début de ce chapitre.

3.1. Introduction

Le but de cette thèse est la classification des sons de la vie courante dans un habitat pour une application de télésurveillance médicale des personnes âgées ou handicapés. Par conséquent, la première étape consiste à construire un corpus de son afin de définir les classes de sons à reconnaitre. Le système de reconnaissance des sons est généralement composé de deux modules : le premier consiste à détecter les sons dans le bruit et les séparer (entrée pour le deuxième module), le deuxième sert à classifier les sons en entrée (figure 3.1).

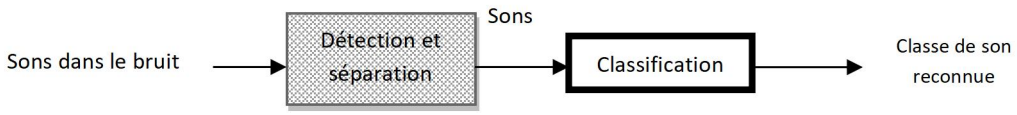

Figure 3. 1. Vue globale d'un système de reconnaissance des sons

En effet, les sons acquis par le système d'acquisition proviennent de différentes sources et sont acquis en même temps. Si nous considérons le nombre de microphones installés dans l'appartement est de n, alors nous avons n sources sonores en un instant t et par conséquent, l'information acquise peut être la même mais avec une variation dans la puissance du signal. Par exemple, le son d'une chute d'objet dans la cuisine peut être acquis simultanément par tous les microphones de l'appartement : salle de bain, salon, etc., mais le son de la chute dans la cuisine est le plus fort et ceci si nous considérons qu'il n'y a pas un bruit dans l'appartement, ou bien plusieurs sons peuvent être acquis simultanément par exemple, son de la télévision dans le salon, son d'écoulement d'eau dans la salle de bain avec un son de toux de l'habitant. Il est donc important d'identifier le son à prendre en compte c.à.d. choisir le son apporteur d'information.

La phase de détection et de séparation est très importante ; qui sera bien détaillée dans le chapitre suivant, elle permet de :

- Éviter la redondance des informations à traiter et donc gagner le temps et l'espace de stockage.
- Séparer les signaux en chevauchement ou bien extraire les sons du bruit de fond.
- Nous ramener à bien déterminer une situation de détresse si elle est réalisée avec précision.

Lorsque les signaux à traiter sont bien choisis, il est temps de les classifier. La classification se fait à deux niveaux :

- Une classification des sons comme sons de la vie courante ou parole,
- La reconnaissance de la classe de son ou de la parole.

Selon la classe de son résultante ou morceau de parole reconnu le système enverra ou non une alarme.

Un système de télésurveillance ne s'agit pas uniquement de la détection des situations de détresse mais il peut servir aussi pour la détection de certaines pathologies, la reconnaissance des activités de la vie quotidienne et le niveau d'autonomie de l'habitant.

Dans cette thèse, nous nous sommes intéressés par le module de classification mais sans oublier de donner une description du module 'détection et séparation' dans le chapitre qui suit.

3.2. Aperçu sur Les sons de la vie courante

Nous désignons par les sons de la vie courante tout type de son qui peut être généré dans l'appartement de l'habitant à savoir la parole, la musique, et autres sons de l'environnement comme les sons de la télévision, sons de vaisselle, claquement, et ouverture ou fermeture de portes.

Les sons environnementaux d'après [Chachada et Kuo, 2013] sont définis par les sons quotidiens (naturels ou artificiels) autres que la musique et la parole. Par conséquent, les sons de la vie courante peuvent être des sons environnementaux, de la parole et de la musique. L'analyse des sons de la vie courante et la classification des différents sons doit donc se faire en tenant compte des trois catégories de sons existantes.

La sélection des sons à étudier pour une application de télésurveillance et d'assistance est une tâche critique car elle dépend de l'objectif global du système à savoir détection de la chute, détection d'une pathologie, (Alzheimer, insuffisance rénale, etc.), détection d'un appel de détresse, etc. Pour cette raison, des travaux antérieurs essayent de rassembler les sons de la vie courante dans des catégories. Istrate [Istrate, 2003], par exemple, divise les sons de la vie courante en deux catégories :

- Sons normaux (reliés aux activités usuelles) tels que le claquement de porte, serrure de porte, sonnerie de téléphone, sons de pas, sons de la vaisselle, etc.
- Sons critiques (possibilité d'existence d'une situation de détresse) comme le bris de verre, chute d'un objet, cris.

Dans [Istrate et al., 2004], 7 classes de sons ont été définies pour une application de télésurveillance : claquement de porte, bris de verre, sonnerie de téléphone, sons de pas, les cris, sons de vaisselle et serrure de porte. Dans [Fleury et al., 2010] et [Vacher, 2011], les auteurs ont défini six catégories de sons excepté la parole : *sons humains* (toux, gargarisme, chant…) qui sont en relation avec la personne, *sons de manipulation d'objets et de fournitures* (manipulation de la chaise, vaisselle, chute d'objet, serrure de porte…) qui sont reliés à l'activité de la personne, *sons extérieurs* (tonnerre, pluie…), *sons de périphériques* (sonnerie de téléphone, bip, tv…), *sons de l'eau courante ;* cette catégorie particulière apporte des informations intéressantes sur les activités telles que l'élimination, l'hygiène, la préparation

des repas. Enfin, la catégorie *autres sons* qui présente tout son qui n'appartient pas aux catégories de sons citées plus haut. Des exemples sur chaque catégorie sont décrits dans le tableau 3.1.

Tableau 3. 1. Catégories de sons dans un habitat

Catégorie de son	Exemples
Sons humains	Toux, gargarisme, soupir, chant, sifflement, essuyage
manipulation d'objets	Fouille d'un sac, Manipulation de la chaise, manipulation de tiroir sons de pas, les chutes d'objets, le bruit du papier
Sons de périphériques	Bip, sonnerie de téléphone, TV
sons d'eau	Lavage des mains, vidange de l'évier, chasse d'eau, écoulement d'eau
Sons extérieurs	Tonnerre, pluie
Autres sons	Bruit provenant d'une source inconnue

Dans le Framework du projet RESIDE-HIS [Castelli et al., 2003], les sons sont divisés en deux catégories : sons utiles (impulsifs et courts) comme la chute d'objets, bris de verre, claquement de porte, et bruit environnemental (longs et stationnaires) comme l'écoulement d'eau, séchoir, rasoir électrique, etc.

3.3. Construction de la base de sons

3.3.1. Description du corpus de sons

La nature de l'application qui est la télésurveillance des personnes âgées ou handicapés exige le type de sons à étudier. Pour cette raison, les classes des sons qui nous semblaient intéressantes ont été définies dans [Abdoune et Fezari, 2016], qui sont divisées en quatre catégories, nous les citons ici : *sons critiques* (cris, chute d'objets, bris de verre, silence en période longue), *sons utiles* (vaisselle, ouverture/ fermeture/ claquement de portes, sons de pas, bâillement, toux), *sons perturbants* (télévision, radio, sonnerie de téléphone, dispositifs électriques, sons extérieurs), parole (*mots clés de détresse* ; j'ai besoin d'aide, j'ai mal Aie ! Au secours, etc.). Le tableau 3.2 ci-dessous présente les sons qui peuvent être générés dans l'habitat et les classes de sons nécessaires pour notre application.

Tableau 3. 2. Sons générés dans un habitat

Son			*Parole*	
Sons critiques	***Sons normaux***		***Parole quotidienne***	***Mots de détresse***
	Sons utiles	**Sons perturbants (bruit)**		
Cri, Chute d'objets, Bris de verre, Silence pour une longue période	Sons de vaisselle, fermeture/ouverture de porte, Claquements de portes, Sons de pas, Écoulement d'eau, Toux, bâillements.	TV, Radio, Sonnerie de téléphone, Sons des dispositifs électriques, Bruit externe.		Besoin d'aide, Au secours, Aïe ! J'ai mal, …

Le principe de notre choix de sons est, d'une part, d'étendre le nombre de classes de sons autant que possible pour assurer une meilleure couverture du domaine d'application et mieux couvrir des scénarios réels, et d'autre part, exploiter cette base de données pour atteindre d'autres objectifs tels que la reconnaissance des activités. Par conséquent, il convient de noter que, même si le son ne semble pas significatif, il est impliqué dans la reconnaissance des activités lorsqu'il est combiné avec d'autres informations.

La base de données utilisée pour notre étude est très diversifiée, une bonne partie de cette base est extraite à partir d'une base de données collaborative en ligne appelée Freesound. Une deuxième partie de cette base est obtenue par enregistrement des différents sons. Les conditions d'enregistrements sont différentes que ce soit au niveau matériel ou logiciel et le nombre d'exemples dans chaque classe est aussi variable.

- Nous avons choisi une fréquence d'échantillonnage de 44,1 KHz afin de restituer fidèlement le signal après numérisation.
- Tous les signaux dans la base de données ont une résolution de 16 bits qui est une bonne résolution temporelle et de bonne qualité, permettant un traitement plus rapide et occupant moins d'espace par rapport au formats avec des résolutions plus élevées.
- Nous avons choisi le format «.wav » pour les fichiers de sons car c'est un format standard et il peut être lu par différents logiciels et sa conversion vers d'autres formats est facile [Istrate, 2003].
- Le rapport signal sur bruit RSB des enregistrements va prendre plusieurs valeurs qui varient entre 0 et 40-60 dB.

Comme il a été expliqué par Dufaux dans sa thèse [Dufaux, 2001], dans la construction d'une BDD, la sélection des signaux influence les résultats d'identification et il faut en tenir compte car les algorithmes de reconnaissance dans leurs principes tentent de placer de manière optimale les limites entre les classes de signaux en fonction d'exemples de ces signaux, et si de mauvais exemples sont donnés au système de reconnaissance au moment de l'apprentissage, les performances seront mauvaises. L'exemple le plus proche est lorsque dans une classe donnée il existe qu'un seul signal qui est différent d'autres sons, le domaine spatial attribué à cette classe sera agrandi, ce qui permettra une plus grande confusion avec certains types de sons voisins. Dans la même classe de sons les sons peuvent avoir différentes caractéristiques temporelles ou spectrales.

Pourquoi Freesound ?

Freesound est une base de données audio qui a été largement utilisée dans différents travaux de recherche sur la reconnaissance de l'audio. Parmi les caractéristiques qui rendent cette base assez importante est le nombre élevé des échantillons audio qu'elle contient qui sont hétérogènes et dépassent 160,000 échantillons audio. Afin de montrer l'utilité de cette base, sa fiabilité et son alignement avec nos objectifs, nous voyons intéressant de présenter dans ce qui suit quelques travaux parmi un nombre important d'applications qui l'utilisent avec une brève description de chaque travail.

La base Freesound a été utilisée par Delgado-Contreras et al. Dans [Delgado-Contreras et al., 2014a] et [Delgado-Contreras et al., 2014 b] pour la classification des endroits par l'utilisation des 'empreintes digitales audio' et deux classifieurs Random forest et SVM. Une autre utilisation de la base Freesound est dans [Chu et al., 2009], pour l'évaluation d'un système de reconnaissance des sons environnementaux pour la compréhension d'une scène. Dans ce travail une deuxième base a été utilisée avec Freesound qui se trouve dans [7]. Aussi, elle a été utilisée dans [Chechik et al., 2008] pour générer un système d'extraction d'information à base du contenu audio. De même, You et Li [You et Li, 2012], ont utilisé la base Freesound dans la reconnaissance des sons environnementaux en utilisant la méthode *TESPAR.* Enfin, Uzkent et al. [Uzkent et al., 2012] utilisent aussi cette même base pour l'évaluation de leur système proposé qui s'agit d'un système de classification des sons environnementaux non-parole.

3.3.2. Les classes de sons

Les sons de la vie courante sont nombreux et diversifiés, et chaque type de son est apporteur d'une information spécifique. La nature de l'application qui est la télésurveillance médicale du sujet âgé et handicapé exige et nécessite une certaine catégorie de sons notamment les sons indiquant une situation de détresse, sons aidant à la localisation de la personne tel que les sons de pas, et sons usuels comme la TV, et sonnerie de téléphone. Nous avons considéré comme importants les sons suivants : Chute d'objets, chute de la personne, bris de verre, écoulement d'eau, son de TV ou radio, sonnerie de téléphone, ouverture et fermeture de portes, son de vaisselle, serrure de porte, toux, cri, gémissement, bâillement, sons de pas, sons

de machines (rasoir, machine à laver, aspirateur, etc.), parole (mots clés de détresse). Tous les fichiers sont des fichiers mono-canal, le nombre de classes est de 15 (tableau 3.3).

Tableau 3. 3. Les classes de son de la vie courante dans un habitat

Classe de sons	Nombre d'échantillons
chute de la personne	23
cri	53
bris de verre	48
chute d'objets	35
écoulement d'eau	67
ouverture /fermeture de portes	105
claquement de porte	53
son de vaisselle	98
serrure de porte	78
sons de pas	85
toux	82
bâillement	37
sonnerie de téléphone	98
son de TV ou radio	55
sons de machines	57

Il est à noter que les sons de machines, TV, radio et téléphone : peuvent être rassemblés dans une seule classe appelée '*sons normaux non significatifs*' ou 'bruit' et on obtient donc 13 classes au lieu de 15.

En effet, nous distinguons dans notre corpus trois catégories de sons notamment : sons indiquant une situation de danger, sons aidant à détecter les activités de l'habitant et sons usuels ou normaux. Dans nos tests nous nous focalisons sur les deux premières catégories, quant à la troisième catégorie (sons de machines et son de TV ou Radio) ses sons peuvent être utilisés comme un bruit de fond en vue de tester le système dans des conditions de bruit environnemental. Les images ci-dessous (figure 3.2 et figure 3.3 montrent quelques captures d'écran de quelques sons :

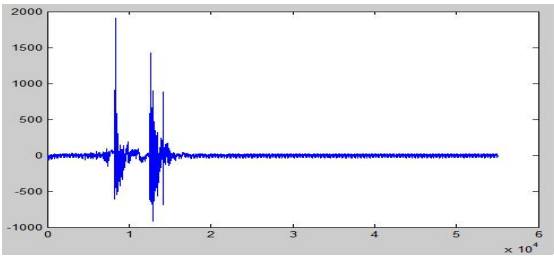

Figure 3. 2. Fermeture de porte

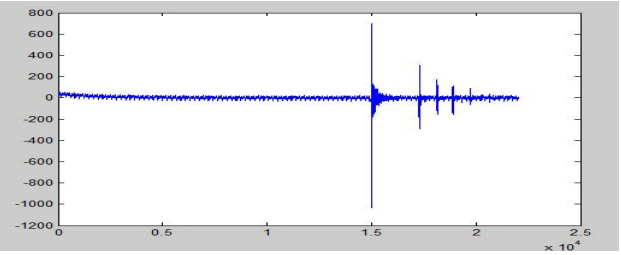

Figure 3. 3. Chute d'objet

Il est pertinent de signaler aussi que pour chaque classe il existe des sons différents qui peuvent provenir de différentes sources, par exemple pour la fermeture de porte on peut trouver plusieurs types de portes avec différentes vitesses de fermeture.

Pour conclure cette partie, la création d'une base de données pour un habitat intelligent est d'une grande importance et en particulier pour les personnes âgées habitant tout seuls où le nombre de sons produits est beaucoup plus petit par rapport au maisons ordinaires. Une description plus détaillée peut se trouver sur [Abdoune et Fezari, 2014] et [Abdoune et Fezari, 2016]. Enfin, malgré que le nombre des échantillons de la présente base de donnée est très petit pour le benchmarking des applications mais elle parait intéressante et peut être étendue par la suite, en augmentant le nombre de sons de chaque classe et en intégrant aussi d'autres versions avec du bruit environnemental pour tester le système dans des conditions bruitées, et en utilisant aussi des techniques d'augmentation des données et surtout pour les sons qui présentent des difficultés lors de leur acquisition et les sons difficiles à reproduire.

3.4. Rapport signal sur bruit

Le rapport signal sur bruit (RSB) sert à mesurer la qualité du signal. Le RSB est le rapport en dB entre la puissance moyenne du signal et celle du bruit [Istrate, 2003]. C'est aussi le rapport, en dB, entre la somme moyennée des carrés des échantillons du signal et la somme moyennée des échantillons du bruit. L'équation (22) donne la formule standard de calcul du RSB.

$$RSB = 10.\log\left(\frac{P_{signal}}{P_{bruit}}\right) = 10.\log\left(\frac{\frac{1}{N}\sum_{i=0}^{N-1} s_i^2}{\frac{1}{M}\sum_{i=0}^{M-1} b_i^2}\right) \quad (22)$$

Où

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

En effet, pour des besoins de tests et d'évaluation et pour une bonne évaluation des performances il est préférable de définir différentes valeurs d'RSB afin de tester les performances de notre système et sa capacité à reconnaître les sons dans des conditions bruitées. Le calcul d'RSB a pour objectif la construction d'un modèle de classe bruité en vue de la construction d'un corpus de son bruité.

Comme décrit dans [Dufaux, 2001] et [Istrate, 2003], il est difficile de trouver la valeur exacte d'RSB, car en supposant que le bruit est généralement stationnaire, alors sa puissance moyenne est constante dans le temps. En revanche, la puissance d'un signal impulsionnel varie rapidement dans le temps. Pour cette raison la mesure d'RSB est difficile et uniquement des approximations peuvent être établies. Une des approximations possibles est le calcul de la valeur instantanée du rapport signal sur bruit. Comme mentionné dans [Istrate, 2003], nous pouvons utiliser une approximation de la puissance moyenne du signal, soit sa valeur

maximale, soit la moyenne sur toute la longueur du son. Dans notre cas, l'RSB est calculé sur toute la longueur du son impulsionnel en tenant compte de la moyenne de l'énergie du signal en rapport avec celle du bruit sur la même durée de la manière suivante (équation 23) :

$$RSB = 10.\log\left(\frac{\sum_{i=0}^{N-1} s_i^2}{\sum_{i=0}^{N-1} b_i^2}\right) \tag{23}$$

Sachant que :

La puissance moyenne du bruit est calculée sur le nombre N d'échantillons du signal et non sur le nombre d'échantillons du bruit comme dans la première équation (équation 22).

3.5. Expérimentation

Une première expérimentation que nous pouvons décrire par préliminaire consiste en la reconnaissance de quelques types de sons avec utilisation d'une simple méthode de classification qui est la distance euclidienne et les caractéristiques ZCR (Zero Crossing Rate) et l'énergie appliqués à la base de sons que nous avons construit nous-même qui est très petite pour pouvoir être testée avec des méthodes de classification et d'extraction de caractéristiques plus compliquées. L'idée donc est de faire ses premiers tests dans le but même de continuer dans la phase de création d'une base de données plus vaste et plus standard que nous pouvons laisser en perspectives et travaux futurs. Cette expérimentation a été intégrée dans un travail qui porte en premier lieu sur la proposition d'un corpus de sons pour la vie courante. Les détails de cette expérimentation se trouvent dans [Abdoune et Fezari, 2016].

Dans cette expérimentation nous avons utilisé une base de données qui contient les sons suivants : les cris, frappe à la porte, sonnerie de téléphone, chute d'objets, sons de vaisselle, adhan et parole. Le format des fichiers audio est .wav et la fréquence d'échantillonnage est 8 khz.

Dans la phase d'extraction de caractéristiques, nous avons utilisé trois paramètres acoustiques : le ZCR et l'énergie en plus des formants qui jouent un rôle important dans la manière dont nous distinguons les différents sons de la parole.

- **Zero crossing rate (ZCR):** indique le nombre de fois le signal dans le domaine temporel traverse l'axe des abscisses x pour un frame donné.

$$\mathrm{ZCR(i)} = \left(\frac{1}{2\mathrm{N}}\right)\sum_{\mathrm{n}=1}^{\mathrm{N}}\left(\left|\mathrm{sign}\left(\mathrm{x_n(i)}\right) - \mathrm{sign}\left(\mathrm{x_{n-1}(i)}\right)\right|\right) \tag{24}$$

- **Short-term energy**

$$E(i) = \left(\frac{1}{Wl}\right)\left(\sum (x_i^2)\right) \tag{25}$$

Où Wl est la longueur de la fenêtre et x_i est l'amplitude du signal à l'instant t

Principe

Après calcul des paramètres acoustiques de tous les sons de la base de données on obtient des vecteurs acoustiques. Le système compare entre les vecteurs acoustiques de chaque segment du fichier test avec les vecteurs acoustiques de la base de données en calculant la différence entre les vecteurs acoustiques de la base de données et les vecteurs acoustiques du fichier test à classifier. Ensuite, par test de vraisemblance le système décide si le segment du son en cours appartient à une classe ou non.

La fonction statistique développée en Matlab classifie le son dans l'une des classes suivantes : frappe à la porte, cris, sonnerie de téléphone et adhan) ; cette fonction utilise la comparaison des paramètres acoustiques et calcule la distance la plus proche.

La figure 3.4 présente l'interface graphique (GUI) de l'application de classification des sons de l'environnement. L'application nous donne des informations sur le son à classifier tel que la durée, fréquence d'échantillonnage, représentation temporelle du signal, spectrogramme, le pourcentage d'appartenance aux différentes classes, et la classification du signal.

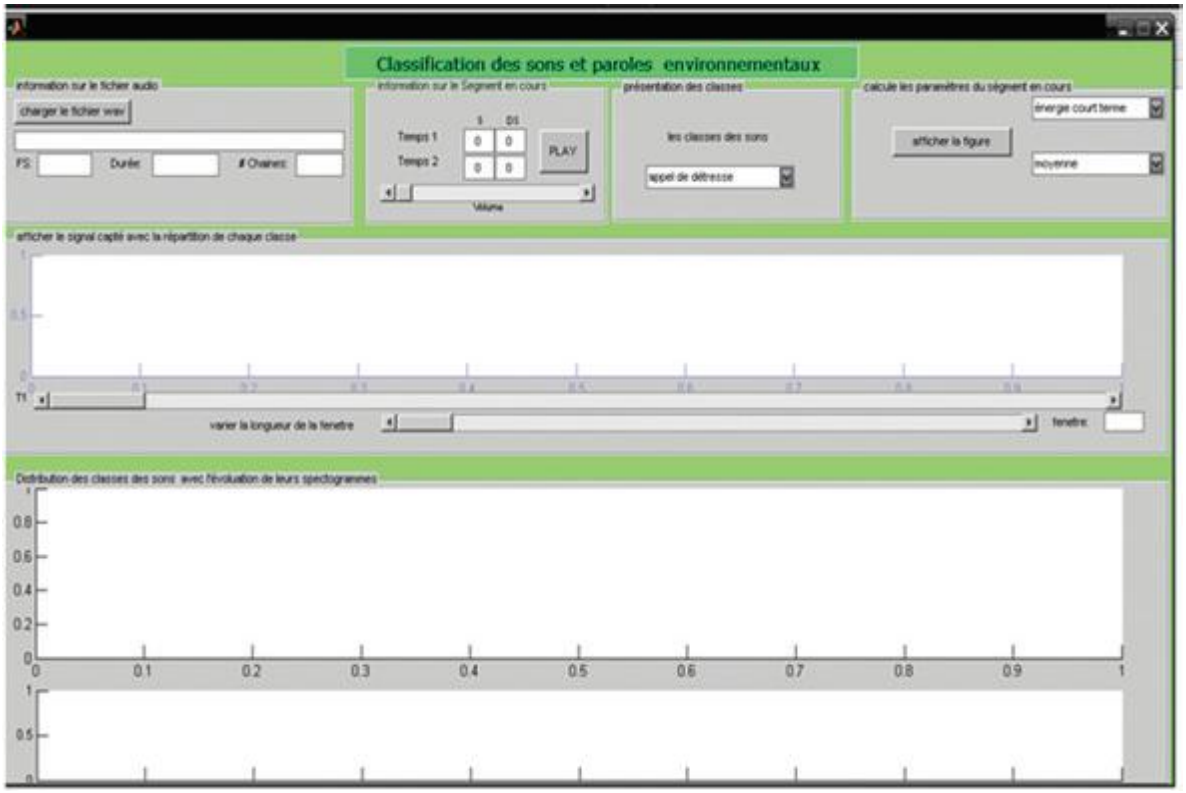

Figure 3. 4. Interface graphique de l'application de classification des sons

Nous avons effectué quelques tests sur 7 types de sons : frappe à la porte, cris et parole, vaisselle, Adhen, ouverture et fermeture de portes, sonnerie de téléphone et chute d'objets.

Les figures (figure 3.5, figure 3.6 et figure 3.7) respectivement, représentent le résultat de classification des sons : frappe à la porte, cris et parole.

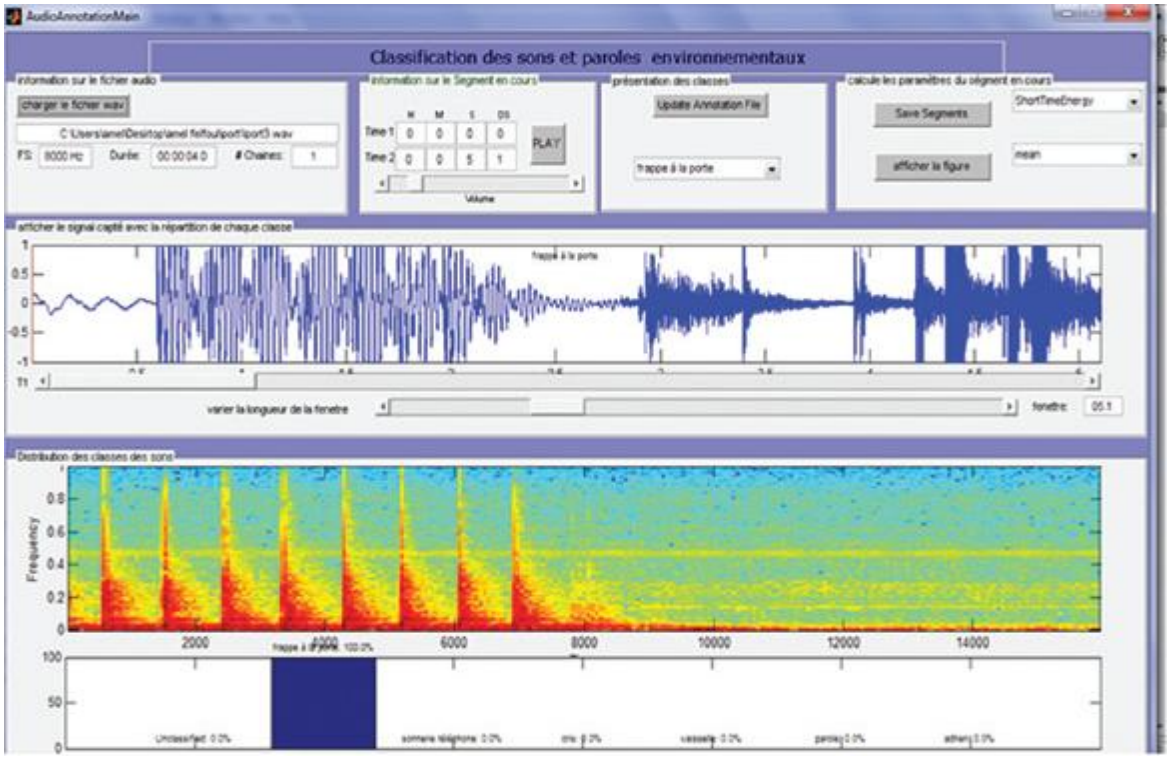

Figure 3. 5. Classification du son frappe à la porte

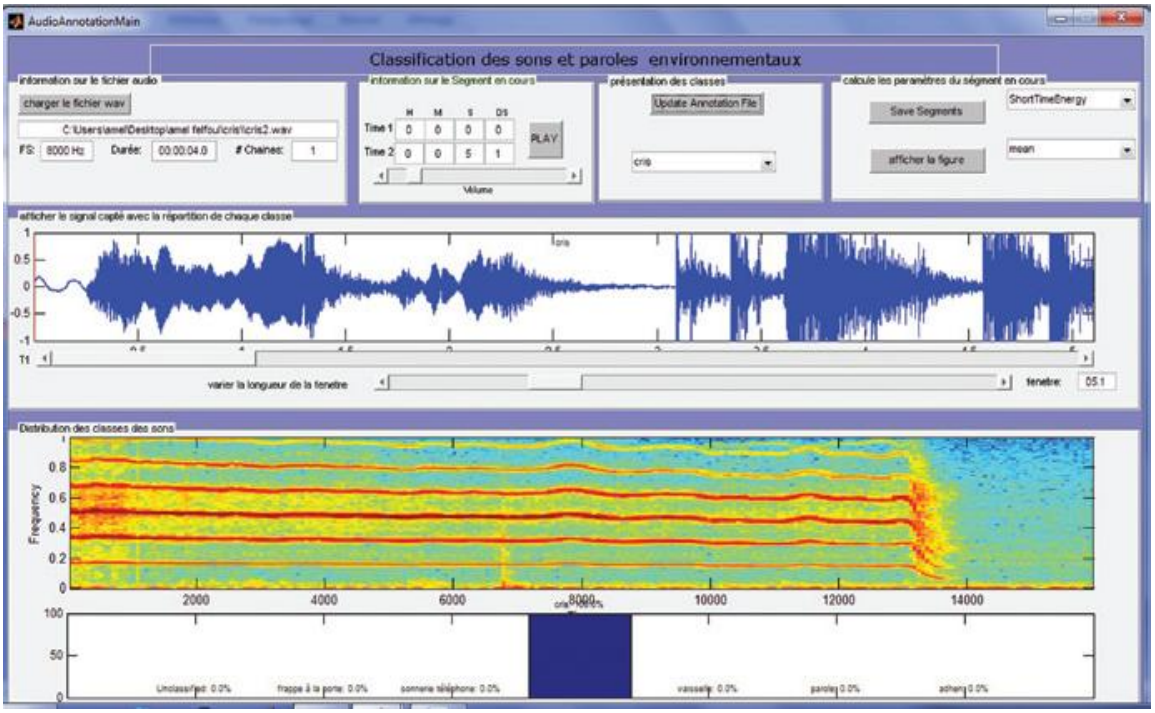

Figure 3. 6. Classification des cris

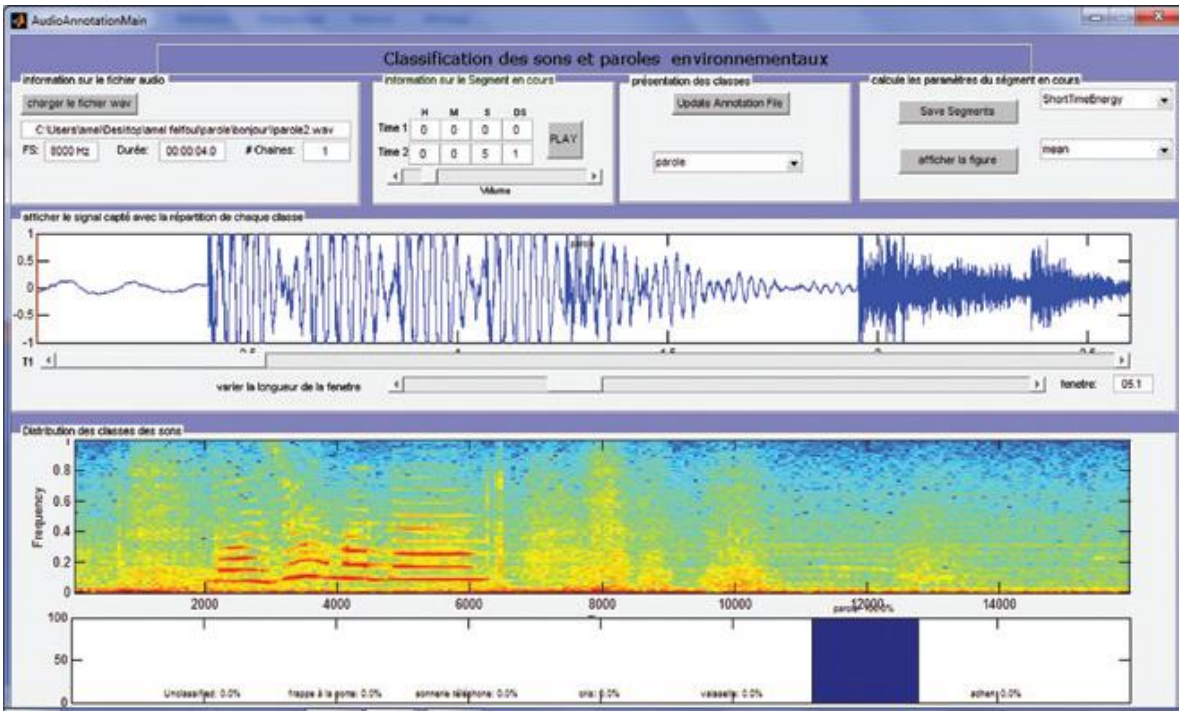

Figure 3. 7. Classification de la parole

Dans cette première expérimentation, nous avons choisi Matlab comme environnement de développement car il offre plusieurs avantages. Il contient une variété d'outils statistiques et de traitement de signal, ce qui permet aux utilisateurs de générer des signaux et de visualiser les graphes. Les résultats de classification pour l'ensemble des classes citées auparavant sont présentés dans la figure 3.8.

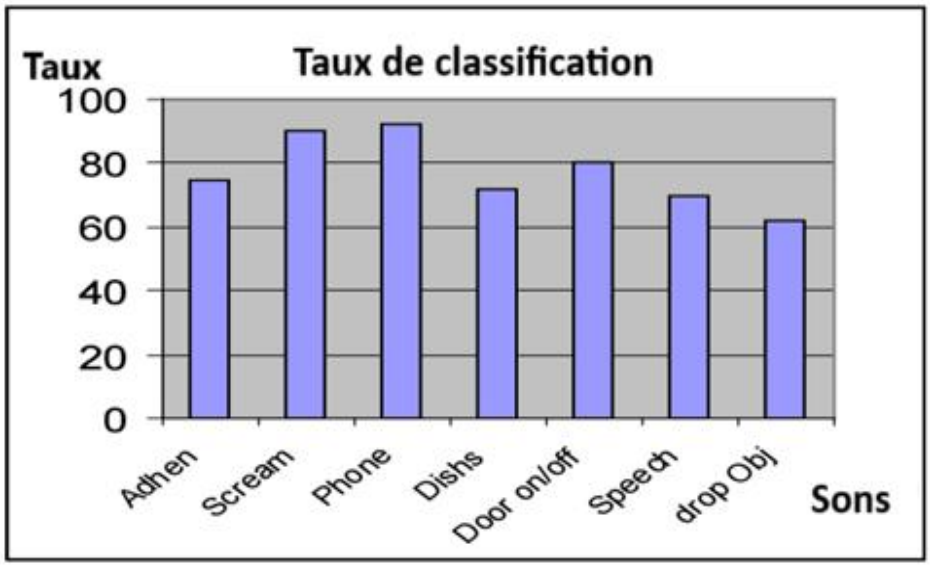

Figure 3. 8. Résultat de classification des 7 classes de son

- **Discussion**

L'objectif de cette expérimentation est de démontrer la nécessité de créer une base de données pour les sons de la vie courante qui était un des objectifs de ce travail au début de notre recherche. Les tests ont été effectués sans aucune étude préalable sur les types de caractéristiques à utiliser et les méthodes de classification à choisir parmi un nombre important de méthodes de classification citées dans la littérature. Nous avons utilisé les trois paramètres acoustiques suivants : le ZCR, l'énergie et les formants. Comme montré dans les figures ci-dessus, les résultats sont encourageants et les différents sons sont reconnus. Les taux de reconnaissance varient de 62% jusqu'à 96% que nous voyons acceptables vu la variation du nombre d'échantillons d'une classe à l'autre et aussi la variation des conditions d'enregistrement. Un facteur important qui a conduit à de faibles taux de reconnaissance est le nombre faible des échantillons par classe.

3.6. Conclusion

Le travail présenté dans ce chapitre porte sur la construction d'un corpus de sons pour la vie courante dans un habitat, dans l'objectif de construire une base de données sonore pour le développement et l'évaluation d'un système de reconnaissance des sons. Des outils simples d'enregistrement ont été utilisés que ce soit logiciels ou matériels et une dizaine d'échantillons a été créée pour chaque classe de sons. Afin de tester cette petite base, les paramètres acoustiques ZCR, énergie et formants sont utilisés, et la classification est effectuée par le test de vraisemblance. En effet, les résultats obtenus sont encourageants, et quoique nous n'ayons pas poursuivi dans cette voie et que nous avons opté dans nos tests pour une plus grande base de données, nous percevons bien l'intérêt d'insérer cette partie dans notre thèse sachant qu'une partie de ce travail a été publiée dans [Abdoune et Fezari, 2016].

Comme perspectives de ce travail nous citons : l'augmentation du nombre de classes et le nombre d'échantillons par classes, et si nécessaire, procéder aux techniques d'augmentation des données et surtout pour les sons difficiles à acquérir tels que la chute de la personne, afin de rendre possible l'application des techniques d'apprentissage profonds.

En résumé, avoir une base de données standard dans le domaine de la reconnaissance des sons environnementaux est indispensable pour le test des applications et surtout pour tirer des conclusions plus précises sur la puissance des algorithmes de reconnaissance des sons et même les méthodes d'extraction de caractéristiques lorsque ces applications sont comparées et utilisent la même base de données.

CHAPITRE 4
Système proposé

Dans ce chapitre, nous décrivons les méthodes que nous avons retenues pour la classification des sons de la vie courante pour la détection d'une situation de détresse qui est l'objet de cette thèse. Avant de ce faire, nous présentons d'abord l'architecture globale du système de télésurveillance, puis les paramètres acoustiques utilisés et enfin, une description détaillée du classifieur est réalisée.

4.1. Introduction

L'identification des classes de sons fait partie du processus de reconnaissance. Nous entendons par identification des classes de sons la classification des différents sons de l'environnement, cependant une étape primordiale pour la classification est la détection d'abord des changements dans l'environnement acoustique, c'est la phase de *détection de son* [Dufaux, 2001]. Après l'étape de détection, vient l'étape de *classification des sons*, qui consiste à identifier le signal détecté. Les deux étapes détection et classification forment le *système de reconnaissance*, or, il est à noter que dans ce manuscrit les deux termes reconnaissance des sons et classification des sons sont utilisés pour désigner la même chose. Dans ce chapitre nous nous focalisons sur l'étape ou le sous système de classification des signaux audio.

Comme décrit dans l'introduction, le contexte de notre travail est la téléassistance des personnes âgées via le canal audio. Le système de reconnaissance des sons générés dans l'habitat peut détecter une éventuelle situation de détresse, comme il peut aussi servir pour la reconnaissance des activités de l'habitant lorsqu'il est combiné avec d'autres systèmes utilisant d'autres capteurs autres que les microphones tels que les caméras, les contacteurs de portes, l'infrarouge, l'accéléromètre, etc.

Ce chapitre étudie la classification des sons de la vie courante. Le corpus de sons utilisé dans l'expérimentation a été présenté dans le chapitre précédant chapitre 3, et les principales raisons de nos choix que ce soit au niveau méthodes d'extraction des descripteurs audio ou méthodes de classification sont décrits dans le chapitre 2.

Avant de s'attaquer aux méthodes mises en œuvre dans ce travail et les résultats obtenus, nous décrivons d'abord l'architecture globale du système de reconnaissance des sons et nous mettons en relief la partie ou le module à réaliser, qui est un système de reconnaissance de sons. Ce module consiste à associer les sons acquis par les microphones à une des classes de sons existantes.

4.2. Architecture générale du système de reconnaissance des sons

L'idée de ce travail est de concevoir un système de reconnaissance de sons pour une application de télésurveillance des personnes âgées vivant seules. Par conséquent, l'application fait aussi partie des systèmes dédiés pour les maisons intelligentes. Pour ce faire, des microphones installés partout dans l'appartement sont utilisés pour capturer les sons qui peuvent se produire à l'intérieur. Le système doit donc être capable d'analyser les sons acquis pour pouvoir les classifier afin de détecter une éventuelle situation de danger dans l'appartement tels que les cris, les appels de détresse et la chute. En effet, l'objectif de cette thèse est en premier lieu l'analyse du domaine de la reconnaissance des sons à travers l'exploration des travaux effectués dans ce domaine et en se focalisant sur les méthodes utilisées, et notre second objectif est de tester les méthodes utilisées pour la reconnaissance de quelques sons, par conséquent, il nous parait important de présenter l'architecture générale du système de reconnaissance de sons qui est

composé de plusieurs modules, mais aussi faire un tour d'horizon sur les solutions offertes pour ce type de systèmes. Pour ce faire, nous avons pensé qu'il serait utile de présenter aussi, critiquer et ainsi que comparer les solutions qui peuvent être adoptées dans chacun des modules de cette architecture et même de montrer les différences clés dans l'architecture elle-même pour chacun des travaux que nous allons présenter.

En supposant, qu'il existe plusieurs microphones dans la maison, le système d'acquisition va prendre en charge uniquement le signal avec l'énergie la plus forte et éliminer le reste. Deux situations, dans ce cas, peuvent se produire :

- Un seul son qui se produit : lorsqu'un seul son est acquis par les microphones donc le système le traite plus facilement selon les architectures proposées et que nous décrivons ci-dessous. Comme on a plusieurs microphones, Le système mesure l'énergie des signaux de tous les canaux et choisit le signal le plus fort qui est celui dont l'énergie est la plus grande.
- Plusieurs sons qui se produisent en même temps ou en chevauchement : dans cette situation un traitement supplémentaire doit être fait sur les sons captés afin de les séparer pour un traitement ultérieur.

A nos connaissances, d'après les travaux faits sur la reconnaissance de sons pour un système de télésurveillance, il existe deux manières principales pour la reconnaissance de sons :

- Soit en intégrant une phase de catégorisation (classifieur binaire) qui permet de distinguer un signal de parole d'un signal de type son, ensuite selon le type de son en sortie (soit son ou parole), le signal est dirigé pour être traité par le système approprié [Vacher et al., 2003] et [Istrate et al., 2008].
- Une deuxième solution consiste à traiter directement le signal en entrée par le classifieur et de considérer la parole comme une classe en plus des classes de sons à reconnaitre tel que le travail de [Wang et al., 2008]. Dans cette deuxième solution, lorsque la classe de son reconnue est celle de la parole un autre système peut servir pour la reconnaissance de la parole en entrée.
- Une autre solution peut être l'intégration de nouvelles classes qui correspondent à la parole en fixant des mots clefs de détresse tels que au secours, de l'aide, etc. Cependant, dans cette solution le temps de réponse du système serait élevé vu l'augmentation du nombre de classes à reconnaitre et par conséquent, augmentation du temps de traitement nécessaire.

Commençons par la solution adoptée par [Wang et al., 2008] qui permet après suppression du bruit du signal en entrée une classification de ce dernier par un classifieur SVM. Le système de classification a 6 classes en sortie y compris la parole (sonnette, bris de verre, frappe, sonnerie de téléphone, toux et parole). Lorsque la classe reconnue correspond à la parole, ce signal est dirigé ou acheminé au sous-système de reconnaissance de la parole pour sa reconnaissance. Le

principe est d'activer certains services d'automatisation selon les classes de sons détectées et s'il s'agit de la parole le système peut détecter les intentions humaines de l'habitant.

Le travail de Istrate dans [Istrate, 2003], traite la détection du son dans un environnement bruité et présente aussi une approche de classification. Le système proposé est composé de deux modules principaux : le module de détection et le module de classification.

- *Le module de détection* : permet une estimation du début et la fin du signal, il permet un gain de temps. Pour la détection 3 algorithmes de détection des ont été testés : détection par corrélation croisée, détection basée sur la prédiction de l'énergie, et détection basée sur le filtrage par ondelettes. Le dernier Algorithme qui est fondé sur la décomposition en ondelettes présente les meilleures performances en présence du bruit réel et un temps de calcul acceptable.
- *Le module de classification* : Pour la classification du son, les modèles de mixtures gaussiennes GMM ont été utilisés.

Istrate dans une autre alternative dans [Istrate, 2003], il propose une approche d'identification des sons clés dans un signal continu en ignorant la parole par recherche de sons clés.

Vacher dans son travail [Vacher, 2011], décrit la phase de détection par la phase d'extraction du signal de parole ou de son du bruit environnemental. Il s'agit de déterminer à chaque instant la présence ou l'absence de signal dans le bruit de fond, pour permettre l'isolement de l'événement sonore qui peut correspondre à un son de la vie courante ou à de la parole, en vue d'un traitement ultérieur. La transformée en ondelettes est une des méthodes de détection les plus efficaces [Vacher et al., 2004]. Après cette phase de détection vient une phase de classification binaire du signal en son ou parole. Enfin, si le signal est parole, il est traité par un classifieur de parole ; le système RAphael [Vaufreydaz et al., 2000] dans son cas, et s'il est de type son il est traité par le sous système de classification des sons de l'environnement.

Dufaux dans sa thèse [Dufaux, 2001], traite la détection et la classification des sons impulsifs. Les méthodes de détection sont basées sur une mesure continue des variations d'énergie du son. Selon le mode du fonctionnement du système de télésurveillance qui est le temps réel par analyse continue du signal de son 24/24, les méthodes de détection doivent être conçues pour représenter une faible charge de calcul, car elles sont destinées à être exécutées tout le temps. Par conséquent, le même principe doit être suivi pour la phase de classification. 10 classes de sons sont traitées dans cette application notamment : claquements de portes, bris de verre, cris humains, explosions, coups de feu, aboiements de chiens, coups bas, sonneries de téléphone, voix d'enfants et machines en marche. Le système est dédié pour les applications de surveillance et de sécurité telles que la détection d'alarmes. Une des raisons pour l'utilisation de cette stratégie qui est la détection puis la classification est la réduction des données à traiter en pensant aux appareils portables dont les capacités sont limitées. Enfin, ce travail ne s'intéresse pas au type du signal détecté (son ou parole), mais plutôt à la classification des signaux en dix classes qui sont fixées à l'avance.

Finalement, Rouas dans son travail [Rouas et al., 2006], aborde le problème de reconnaissance des sons pour une application de télésurveillance dans un transport public. Le son à détecter est les cris. Il propose dans son système une étape de prétraitement considérée comme première étape du système, elle permet la détection et la segmentation du signal en entrée afin de conserver uniquement les zones d'intérêt du signal et éliminer le reste. Ceci est fait en trois étapes :

- *Segmentation audio automatique* qui consiste à décomposer le signal en un ensemble de segments. La segmentation est faite par l'utilisation de l'algorithme DFB (Forward-Backward Divergence).
- *Détection d'activités* en supprimant les zones représentant le silence et les zones de bruit.
- *La fusion* qui consiste à rassembler les segments d'activités successives.

Ensuite, le signal en sortie est classifié soit comme parole ou non parole, pour être enfin classifié par un classifieur binaire ; soit cris ou non cris.

Le tableau 4.1 ci-dessous, résume l'ensemble des architectures des systèmes de reconnaissance des sons présentés plus haut. Nous nous focalisons sur les modules contenus dans les différentes architectures ainsi que les éventuelles méthodes de détection et de segmentation, en plus des types de sons à classifier. Les méthodes de classification ne nous intéressent pas ici étant donné qu'elles ont déjà été abordées dans les chapitres précédents.

Tableau 4. 1. Architectures des systèmes de reconnaissance des sons

Architecture	Application	Modules	Méthodes	Sons à classifier
[Dufaux, 2001]	Détection et la classification des sons impulsifs pour la surveillance et la sécurité	-Détection -Classification	Variations d'énergie pour la détection	10 classes de sons (indoor et outdoor)
[Istrate, 2003]	Détection et reconnaissance des sons pour la surveillance médicale	-Détection -Classification par GMM	Test de 3 méthodes : Corrélation croisée, prédiction d'énergie et filtrage par ondelettes	Sons produits dans un habitat
[Rouas et al., 2006]	reconnaissance des sons pour une application de télésurveillance dans un transport public	- Segmentation et détection du signal (enlever bruit+ zones de silence) -Classification en parole ou non parole	- Segmentation par l'algorithme DFB - Détection par un algorithme basé sur une analyse statistique du 1er ordre du signal temporel.	cris ou non cris

[Vacher, 2011]	Analyse sonore et multimodale dans le domaine de l'assistance à domicile	- Détection par extraction du son du bruit de fond - Classification binaire (son ou parole) - classification de chaque catégorie de son à part	- Transformée en ondelettes pour la détection	Parole et sons
[Wang et al., 2008]	Reconnaissance des sons environnementaux pour la domotique	-Suppression du bruit - classification (classes de sons + classe parole) - classification de la parole	- Banc de filtres perceptuels + méthode basée sur des sous-espaces pour la suppression bruit	Parole et sons

- **Discussion**

A partir de ces travaux et architectures existantes des systèmes de reconnaissance de sons pour des applications de télésurveillance, nous identifions les diverses solutions adoptées dans chacun des travaux. Nous voyons donc que la solution qui adopte un classifieur binaire pour séparer les deux catégories de sons existantes (parole et son) est la plus réaliste, mais elle doit être faite avec précaution car ses résultats vont influencer la totalité du système. De même, pour la phase de détection des sons, c'est une phase très critique car un signal existant mais non détecté par ce module vas mettre le système entier en risque, et de l'autre côté, ce module va apporter des gains de temps et de traitement en réduisant le nombre d'échantillons à traiter par élimination des signaux présentant des périodes de silence, et même les signaux à faible énergie. Par conséquent, l'ensemble des solutions proposées dans les architectures présentées, après étude et comparaison, nous mènent vers l'architecture illustrée dans la figure 4.1, qui constitue pour nous la solution la plus appropriée et montre une vue globale du système de reconnaissance des sons.

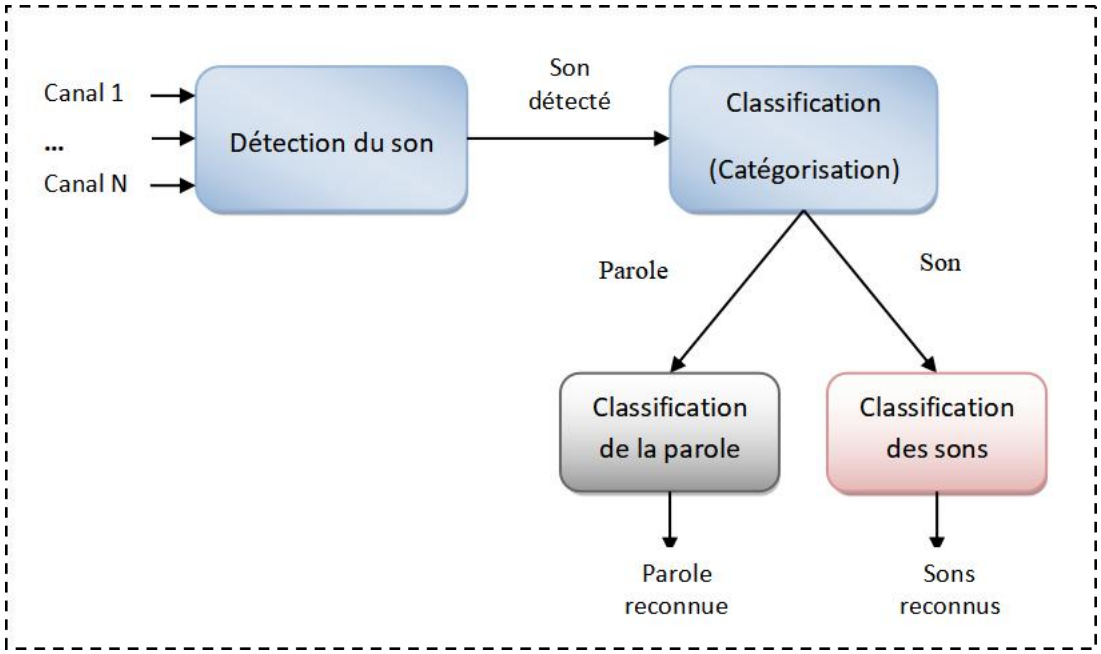

Figure 4. 1. Architecture générale du système de reconnaissance des sons

Le système de télésurveillance est composé de plusieurs modules, en voici une description abrégée des différents modules de l'architecture :

1. Détection de son

La première étape consiste à ***détecter le son*** à travers l'analyse continue d'un flux de données audio en définissant le début et la fin de ce signal. L'algorithme qu'on doit utiliser doit être robuste au bruit vu la nature de l'environnement *maison* où divers sons de type bruit sont présents tels que les sons de machines à laver, séchoir électrique et tendeuse. En plus de ce bruit, il existe le bruit hors maison tel que le tonnerre, les véhicules et la pluie. Dans [Istrate et al., 2008], par exemple, un algorithme à base des ondelettes a été utilisé. Le rôle de la phase de détection des sons est de déterminer l'instant d'apparition d'un événement sonore, en vue de l'extraire du bruit de fond pour qu'il passe à l'étape de classification [Istrate, 2003] et [Rouas et al., 2006]. Cette étape nous permet par conséquent, de réduire le temps de calcul mais aussi améliorer les performances du système en réduisant les informations inutiles telles que le silence, les sons à faible fréquence. En effet, dans la phase de détection, il existe plusieurs sous modules que l'on peut décrire ci-dessous :

- ***Segmentation audio automatique*** qui consiste à décomposer le signal en un ensemble de segments. Il existe plusieurs algorithmes de segmentation tels que DFB (Forward-Backward Divergence) [André-Obrecht, 1988], qui est basé sur l'étude statistique du signal où chaque segment est caractérisé par un modèle statistique.
- ***Détection d'activités*** en supprimant les zones représentant le silence et les zones de bruit.
- ***La fusion*** *qui consiste à rassembler les segments d'activités successives* (voir Figure 4.2).

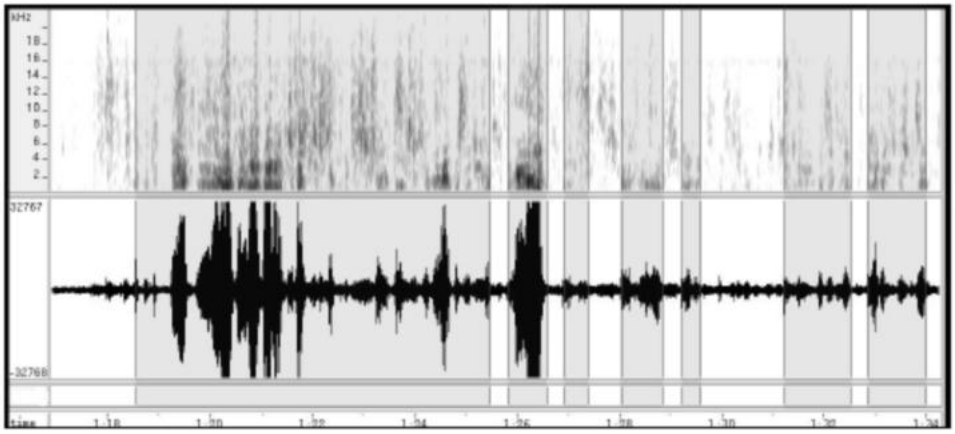

Figure 4. 2. Zones d'activités détectées dans un signal audio (en gris) et zones écartées et supprimées (en blanc [Rouas et al., 2006])

2. Classification du son comme son ou parole

Appelée aussi *phase de catégorisation* : cette étape consiste à classifier le signal audio comme son ou parole, c'est une phase de segmentation du signal en entrée où un classifieur binaire doit être utilisé. Dans [Istrate et al., 2008] un classifieur GMM a été utilisé avec LFCC à 16 filtres et 24 modèles gaussiens (Bayesian Information Criterion (BIC) a été utilisé pour déterminer le nombre optimal de modèles gaussiens à utiliser).

- **Reconnaissance de la parole :** dans cette phase un modèle de langage doit être utilisé, qui sera ensuite optimisé pour les phrases ou mots de détresse. Une référence importante est le système RAPHAEL [Vacher et al., 2010a].
- **Reconnaissance de son :** dans cette étape se fait la reconnaissance des sons en entrée résultants de l'étape précédente.

En effet, c'est dans le sous module reconnaissance de son, que notre travail tourne. Le traitement commence par l'extraction des caractéristiques et se termine lors de la reconnaissance des sons. Par conséquent, la section suivante est consacrée pour une description détaillée du module de reconnaissance de son.

4.3. Architecture du sous système de classification des sons

Dans cette section nous présentons l'architecture du classifieur des sons en mettant l'accent sur les paramètres acoustiques utilisés et la méthode de classification.

Différents travaux ont été effectués sur la RSE qui ont exploré diverses méthodes de classification tels que les GMM et les réseaux de neurones. Notre première motivation fut donc d'explorer d'autres techniques et en particulier les machines à vecteurs support SVM vu leur intérêt.

Dans le chapitre précédent (chapitre 3) nous avons défini les classes de sons à reconnaître qui sont : chute d'objets, chute de la personne, bris de verre, toux, bâillement, écoulement d'eau,

ouverture/fermeture/claquement de portes, son de vaisselle, serrure de porte, sonnerie de téléphone, son de TV ou radio, sons de pas et sons de machines.

Dans cette section, nous décrivons les différentes étapes nécessaires pour le processus de reconnaissance comme indiqué dans la figure ci-dessous (Figure 4.3).

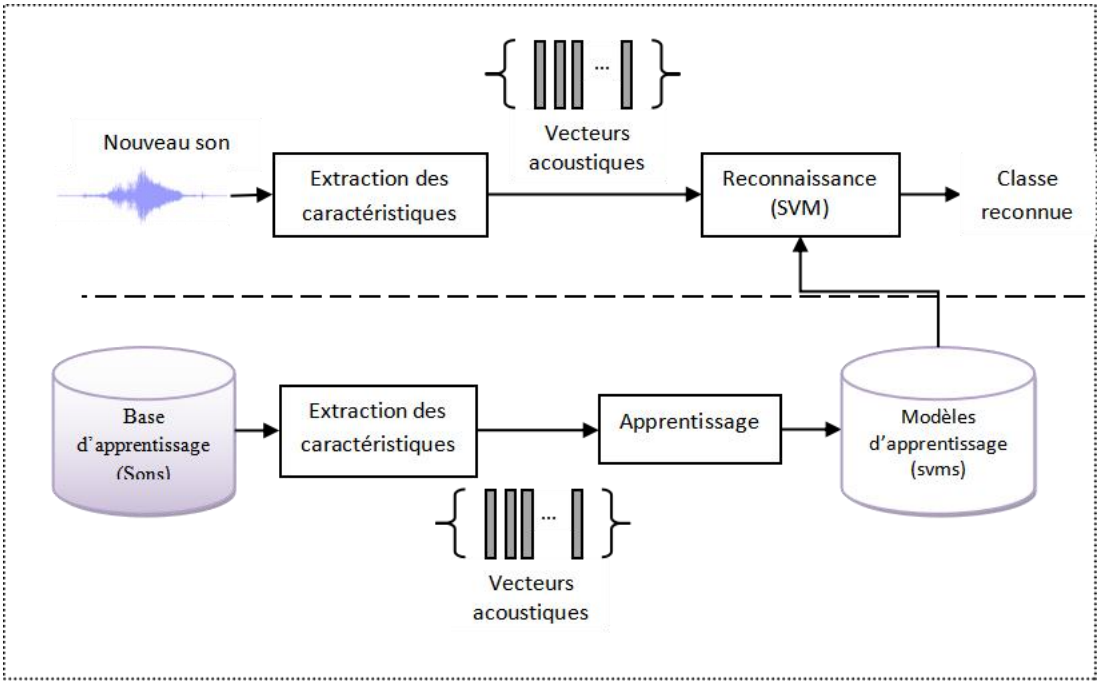

Figure 4. 3. Architecture détaillée du système de classification des sons

4.3.1. L'extraction des caractéristiques

Avant de parler de la phase d'extraction de caractéristiques, des prétraitements peuvent être faits avant et après cette phase. Un premier prétraitement peut être l'élimination du bruit mais vu que les sons sont extraits du bruit de fond dans la phase de détection, nous ignorons cette opération ici. Vient maintenant l'extraction de caractéristiques proprement dite, dans cette phase et comme expliqué déjà dans le chapitre 1, les MFCC sont nos premiers paramètres à utiliser avec le classifieur basé SVM. Pour cette raison, dans ce qui suit nous allons nous limiter à la description de l'extraction des caractéristiques des MFCC.

- **Description du calcul MFCC**

Les MFCC sont calculés après transformation du signal dans le domaine spectral, comme déjà décrit en détail dans le chapitre 1. Dans un premier cas, nous prenons les 12 ou 13 MFCC, comme entrée de notre classifieur, ensuite un **deuxième** test consiste à ajouter à ces coefficients leurs dérivées première et deuxième. Le même traitement qu'a avec 13 MFCC est répété en prenant 40 MFCC.

- **Le fenêtrage ou choix de la longueur de la fenêtre**

Vu la nature des sons à traiter dans cette application qui sont des sons non-parole : très courts, non stationnaires et qui changent de caractéristiques acoustiques très rapidement dans le temps, la longueur de la fenêtre doit être petite. Nous avons appliqué une fenêtre rectangulaire de 20-30 millisecondes avec 50% de chevauchement pour chaque son afin de calculer les paramètres acoustiques. Ces petites unités sont appelées segments, et c'est sur ces dernières que l'extraction des caractéristiques est réalisée. Il est à noter que le choix de la fenêtre à appliquer sur le signal est une tâche critique et dépend des caractéristiques du signal à traiter, car pour un signal de petite durée la taille de la fenêtre doit être petite et suffisante pour capturer les détails des signaux de courte durée et qui changent de caractéristiques rapidement dans le temps, mais pas trop petite sinon on risque d'avoir une mauvaise résolution de la fréquence du signal en question. Inversement, choisir des fenêtres de grandes tailles peut avoir pour effet une mauvaise résolution temporelle, donc une difficulté de capturer les événements qui se produisent dans les petites durées. En résumé, le choix de la fenêtre du signal doit se faire avec prudence, et si nécessaire tester des fenêtres avec différentes tailles et comparer les résultats pour choisir la fenêtre optimale. Des informations pertinentes et intéressantes sur le choix de la fenêtre des signaux peut se trouver sur l'étude de Chachada et Kuo sur la reconnaissance des sons de l'environnement dans [Chachada et Kuo,2014].

- **La normalisation des paramètres**

C'est une étape qui est liée à la méthode de classification utilisée. Les SVM par exemple, nécessitent une étape de normalisation des paramètres. Le principe de la normalisation et les différentes techniques utilisées sont bien décrites dans le chapitre 1.

- **La sélection des paramètres**

Vu que les MFCC sont nos premiers paramètres à tester dans notre travail, la méthode de sélection des paramètres n'a pas été utilisée. Cependant, la combinaison des paramètres acoustiques reste comme un deuxième objectif mais qui est laissé en perspectives. Par conséquent, l'utilisation de la méthode de sélection des paramètres dans le deuxième cas reste indispensable.

4.3.2. Description du classifieur basé SVM

- **Noyaux de description**

Dans notre expérimentation, nous avons essayé d'utiliser le classifieur SVM avec différents noyaux : *le noyau gaussien (RBF)*, *le noyau polynomial, le noyau sigmoïde* et *le noyau linéaire* dans l'objectif de les comparer.

- **Un contre un ou un contre tous ?**

Les deux techniques un contre un et un contre tous sont très utilisées pour passer d'un SVM binaire à un SVM multi-classes, cependant beaucoup de travaux encouragent l'utilisation de la méthode un contre tous qui est performante et moins coûteuse mais avec un temps d'apprentissage plus long. Pour cette raison, nous nous sommes tournés vers l'utilisation de la

méthode un contre tous. Il est indispensable de noter, que le choix entre ces deux méthodes n'est pas aussi facile car plusieurs critères peuvent intervenir en plus du temps de calcul et les performances obtenues tels que la nature du problème à traiter, les caractéristiques de la base de données (petite ou grande taille), le nombre de classes, le fait d'avoir des caractéristiques en chevauchement entre deux ou plusieurs classes, etc. Ces facteurs jouent un rôle important dans le choix de l'une des méthodes dont il faut tenir compte dans toute application. Pour plus de détails, sue le choix d'une méthode ou de l'autre, le travail cité dans [Hsu et Lin, 2002] contient plus d'explications.

- **Corpus de son et base de données**

Nous avons montré dans le *chapitre 3* le corpus de sons de la vie courante, mais l'expérimentation s'est concentrée sur quelques types particuliers de sons à savoir ***les cris, sons de vaisselle, et le bris de verre***. Ensuite, d'autres tests ont été effectués en augmentant le nombre de classes à 7 classes au lieu de 3.

- **Base d'apprentissage et de test**

Il existe différentes façons pour définir la base d'apprentissage et la base de test tel qu'expliqué dans le chapitre 1. Dans notre travail, nous avons suit une décomposition de la base d'apprentissage en 20% pour le test et le reste pour l'apprentissage.

- **L'ajustement des hyperparamètres**

La valeur optimale du paramètre C est fixée à l'aide d'une méthode très classique appelée *grid search*, et il en va de même pour le degré, le gamma et le $coef0$. En effet, cette méthode consiste d'abord à choisir les valeurs des hyperparamètres que l'on souhaite tester, puis on choisit le critère d'évaluation de la qualité du modèle (précision par exemple), après on définit la méthode de validation et on teste toutes les combinaisons. Nous avons appliqué la technique de validation croisée pour obtenir les meilleures valeurs qui optimisent les modèles.

Il est important de noter que les formules associées à chacun de ces noyaux sont bien décrites dans le chapitre 1. De plus, le paramètre C est un paramètre de régularisation commun à tous les noyaux et sa valeur permet de définir la taille de la marge et les erreurs de classification.

- **Résultats**

Dans cette expérimentation nous avons opté pour deux cas d'étude de notre corpus de son :

- Le premier consiste à appliquer notre SVM sur 3 classes de sons notamment, les cris, bris de verres, et les sons de vaisselle.
- Le deuxième consiste à augmenter le nombre de classes à reconnaître en un nombre de 7 classes au lieu de 3, afin de comparer les résultats et de mieux comprendre le comportement du système pour un nombre élevé de classes.

a) **3 classes** (cris, sons de vaisselle et bris de verre)

Pour commencer notre expérimentation, nous avons démarré par le test des caractéristiques MFCC avec deux valeurs différentes respectivement 13 et 40. Sachant que les 12 ou 13 premiers MFCC sont les plus porteurs d'information en reconnaissance de la parole, et dans les sons environnementaux on peut aller jusqu'à 40 descripteurs selon le type du signal traité et son niveau de complexité. Les métriques choisies sont la précision et le F1 score. Les deux tableaux (tableau 4.2 et tableau 4.3) représentent respectivement les performances obtenues par les différents noyaux de l'SVM pour 13 et 40 MFCC et ceci pour 3 classes de sons de la vie courante (cris, sons de vaisselle et bris de verre). Une comparaison des performances obtenues par les différents noyaux est présentée par le graphique de la figure 4.4.

Tableau 4. 2. Taux de reconnaissance du classifieur SVM pour les 4 noyaux (3 classes) avec 13 MFCC

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	85.19%	59.26%	96.30%	100%
F1 score	83.03%	59.61%	96.28%	100%

Tableau 4. 3. Taux de reconnaissance du classifieur SVM pour les 4 noyaux (3 classes) avec 40 MFCC

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	85.19%	59.26%	96.30%	100.00%
F1 score	83.03%	59.61%	96.28%	100.00%

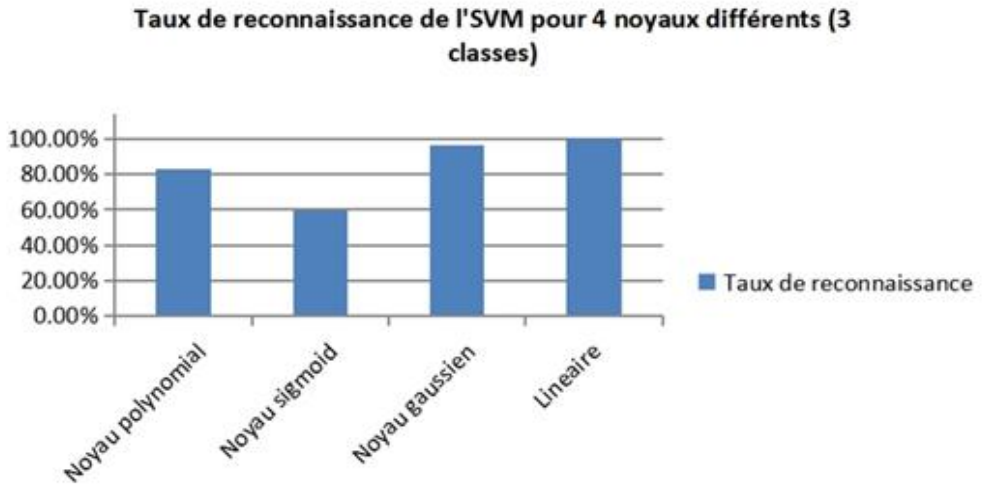

Figure 4. 4. Taux de reconnaissance du classifieur SVM pour des noyaux différents : 3 classes

Nous observons premièrement que Les deux tableaux présentent les mêmes valeurs. Donc, dans le cas de 3 classes l'augmentation du nombre de descripteurs n'a aucun effet sur les performances du système. Ensuite, selon les valeurs obtenues par chacun des noyaux nous remarquons que le noyaux linéaire présente les meilleures performances qui est de 100%,

ensuite le noyau RBF de 96,30% puis le noyau polynomial de 85.19% et de même pour le F1-score. Par contre le noyau sigmoïde présente de mauvaises performances malgré le nombre limité de classes.

A l'issue de cette étude, nous constatons qu'un taux de reconnaissance de 100% pour un SVM à noyau linéaire veut dire que les données sont linéairement séparables et donc les caractéristiques MFCC sont bien discriminantes. En revanche, pour les autres noyaux le taux varie de 96-59%, ce qui est expliqué par le choix arbitraire des hyperparamètres spécifiques à chacun de ces noyaux comme le paramètre C, gamma, et degré.

Les deux figures 4.5 et 4.6 ci-dessous, montrent respectivement la matrice de confusion pour les 3 classes (cris, sons de vaisselle et bris de verre), pour le noyau linéaire et RBF.

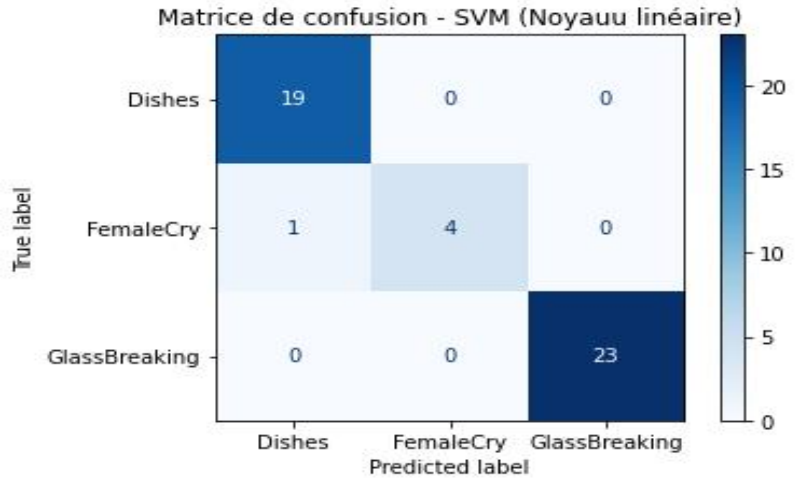

Figure 4. 5. Matrice de confusion pour l'SVM à noyau Linéaire

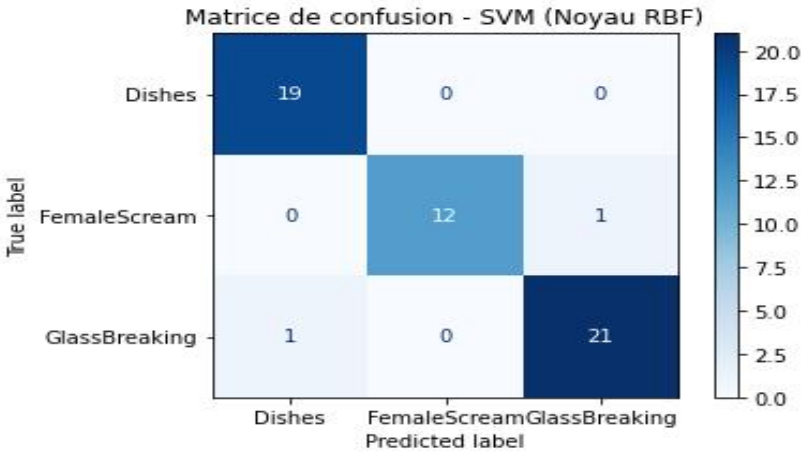

Figure 4. 6. Matrice de confusion pour l'SVM à noyau RBF

b) **7 classes** (bris de verre, ouverture de porte, claquement de porte, cris, vaisselle, toux, écoulement d'eau)

Dans ce test, nous augmentons le nombre de classes à reconnaître en un nombre de 7 classes au lieu de 3, qui sont : bris de verre, ouverture de porte, claquement de porte, cris, vaisselle, toux, et

écoulement d'eau. Les tableaux 4.4 et 4.5 suivants, présentent le f1 score et la précision obtenus par les différents noyaux pour un nombre d'MFCC de 13 puis avec 40 MFCC, respectivement.

Tableau 4. 4. Performances de l'SVM pour les différents noyaux avec 13 MFCC pour 7 classes

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	72.12%	36.54%	70.19%	91.35%
F1 score	67.85%	33.9%	63.57%	91.34%

Tableau 4. 5. Performances de l'SVM pour les différents noyaux avec 40 MFCC pour 7 classes

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	72.12%	36.54%	74.04%	95.19%
F1 score	67.85%	34.35%	68.96%	95.11%

Pour le cas de 7 classes, on observe une diminution des taux de reconnaissance pour tous les noyaux, que ce soit pour 13 ou 40 MFCC. L'ordre de classement des noyaux SVM, du plus performant au moins performant, reste toujours le même ; le noyau linéaire en premier et le noyau sigmoïde est le moins performant. Cependant, dans le cas où le nombre d'MFCC est 40, nous voyons une amélioration significative des performances pour les deux noyaux linéaire et gaussien par rapport aux valeurs obtenues en utilisant uniquement 13 MFCC. La figure 4.7, illustre le graphique qui représente les taux de reconnaissance de l'SVM pour les différents noyaux avec 40 MFCC.

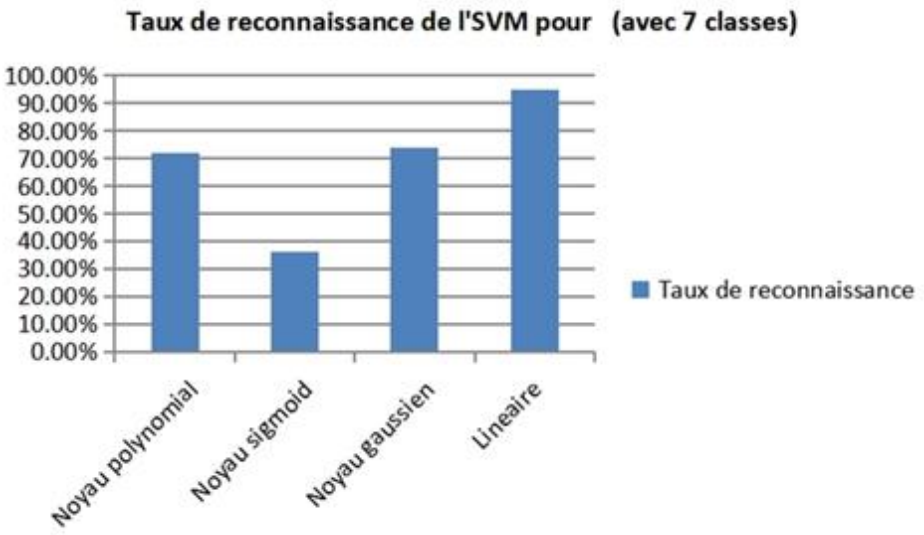

Figure 4. 7. Taux de reconnaissance de l'SVM (7 classes) avec 40 MFCC

c) 7 classes et 40 MFCC avec validation croisée

Dans ce test et les tests qui suivent, nous travaillons sur le cas de 7 classes qui parait plus crédible. Notre expérimentation consiste à utiliser la méthode de validation croisée ou grid search pour l'ajustement des hyperparamètres des différents noyaux. Le tableau 4.6 suivant, montre les taux de reconnaissance ainsi que le F Score obtenus par les quatre noyaux d'SVM, avec comme caractéristiques 40 MFCC.

Tableau 4. 6. Performances de l'SVM pour les différents noyaux avec 40 MFCC après ajustement des hyperparamètres des noyaux.

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	90.38%	36.54%	96.15%	%95.19
F1 score	90.10%	34.35%	96.12%	%95.11
Hyperparamètres	C=100, degree : 2	C=1 gamma: scale	C=100 gamma: scale	C=0.1

Les résultats montrent une amélioration significative des performances des deux noyaux polynomial (90.38% au lieu de 72.12%) et le noyau RBF (96.15% au lieu de 74.04%), ce qui justifie l'utilité de la méthode de validation croisée ou grid search dans l'ajustement des paramètres des noyaux SVM. L'SVM avec un noyau RBF présente les meilleures performances, ensuite l'SVM à noyau linéaire, puis l'SVM à noyau polynomial et enfin, l'SVM à noyau sigmoïde qui se révèle le moins performant (36.54%) en comparaison avec les autres noyaux.

d) 7 classes et 40 MFCC avec normalisation et validation croisée

Après normalisation des données qui est une étape primordiale pour certains types de classifieurs tel que les SVM, nous observons, comme montré dans le tableaux 4.7, une amélioration des performances pour tous les noyaux sauf le noyau linéaire qui a gardé ses performances précédentes. Le noyau RBF s'avère toujours le noyau le plus performant avec un taux de reconnaissance de 97.12% puis le noyau linéaire. Les deux noyaux polynomial et Sigmoïde donnent des performances similaires. La normalisation des caractéristiques a nettement augmenté les performances du noyau sigmoïde (de 36.54% à 92.31%).

Tableau 4. 7. Performances de l'SVM pour les différents noyaux avec 40 MFCC après ajustement des hyperparamètres et normalisation.

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	92.31%	92.31%	**97.12%**	94.23%
F1 score	92.24%	92.20%	**97.11%**	94.11%
Hyperparamètres	C=100, degree= 2	C=1, gamma :auto	C=10, gamma=0.001	C=0.1

La figure 4.8 illustre une comparaison des performances obtenues par les différentes noyaux SVM après une phase de normalisation des caractéristiques, et la figure 4.9 présente la matrice de confusion pour l'SVM à noyau RBF.

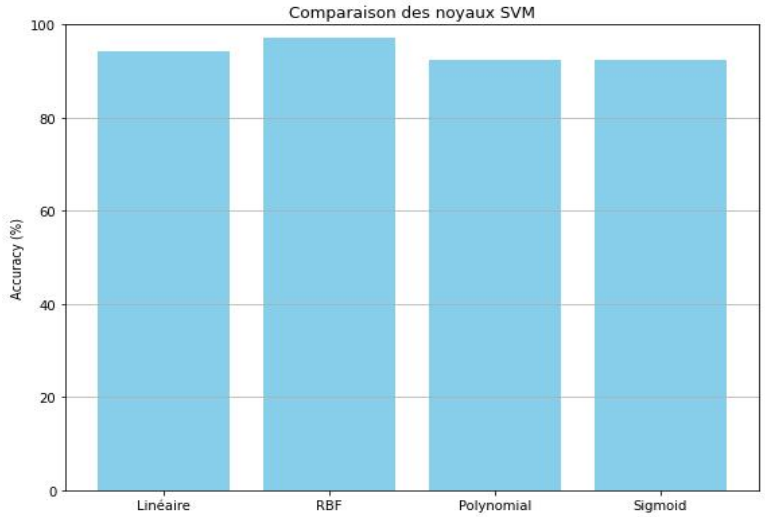

Figure 4. 8. Performances des Noyaux SVM après normalisation

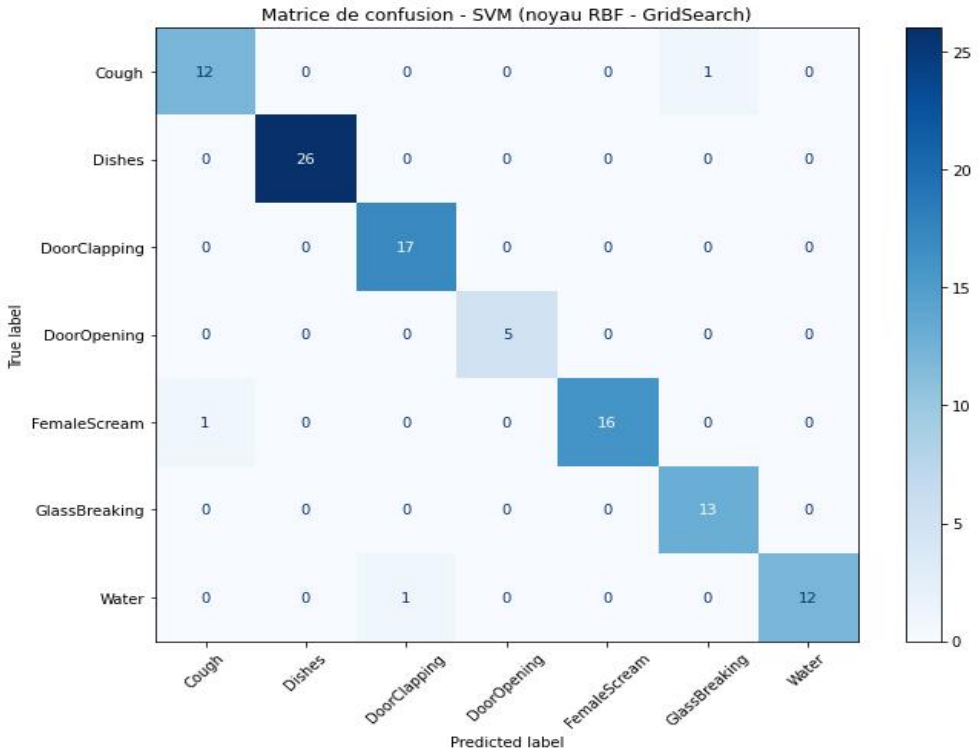

Figure 4. 9 Matrice de confusion pour l'SVM à noyau RBF avec 40 MFCC

e) 7 classes et 13 MFCC avec normalisation et grid search

Avec 13 MFCC et normalisation nous constatons une augmentation des performances du noyau linéaire qui devient similaire au noyau RBF. Une légère augmentation du taux de reconnaissance

offert par le noyau polynomial. Cependant les performances du noyau sigmoïde se dégradent de nouveau de 92.31% à 77.88%. Enfin, le noyau RBF et linéaire présentent les meilleurs résultats. Le tableau 4.8 ci-dessous résume les taux de précision et le F1 score des différents noyaux ainsi que les valeurs prises par les hyperparamètres, et les figures 4.10 et 4.11 présentent respectivement les matrices de confusion pour les deux noyaux linéaire et RBF.

Tableau 4. 8. Performances des noyaux SVM avec les différents noyaux après normalisation et ajustement des hyperparamètres.

	Noyau Polynomial	**Noyau Sigmoïde**	**Noyau RBF (Gaussien)**	**Linéaire**
Précision	94.23%	77.88%	**97.12%**	**97.12%**
F1 score	94.17%	75.85%	**97.07%**	**97.09%**
Hyperparamètres	C: 100, degree: 3	'C': 1, 'gamma': 'auto'	C=10, gamma: scale	C=1

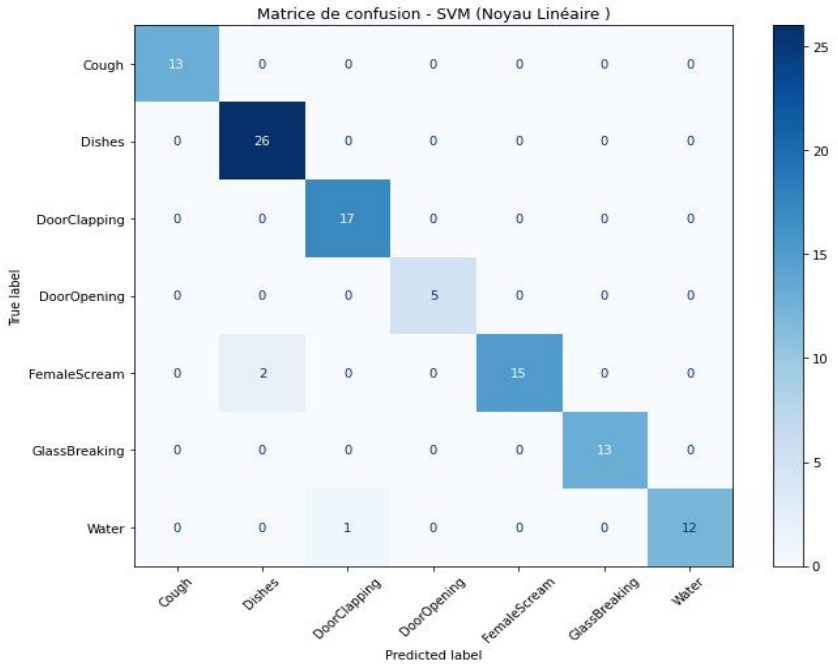

Figure 4. 10. Matrice de confusion pour SVM à noyau linéaire avec 13 MFCC

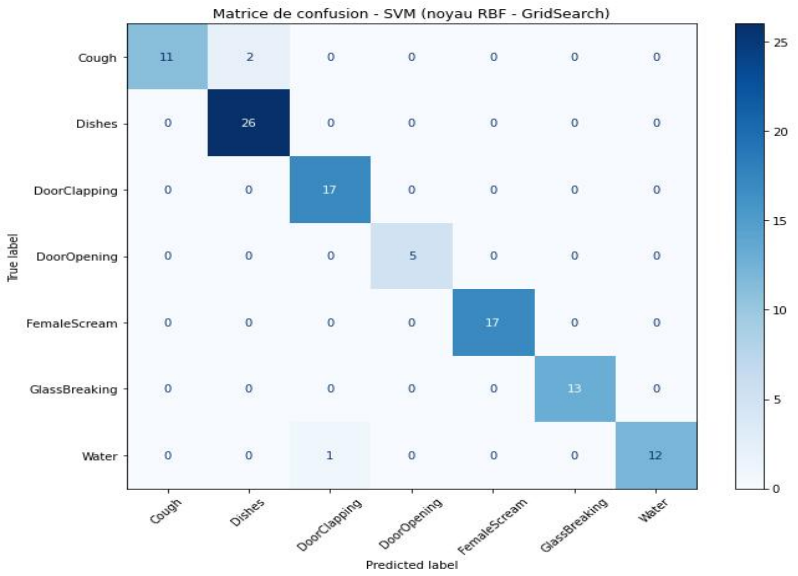

Figure 4. 11. Matrice de confusion pour SVM à noyau RBF avec 13 MFCC

f) 13 MFCC et delta et delta delta MFCC

Après avoir testé les paramètres MFCC avec différentes variantes, nous faisons un nouveau test en combinant les paramètres MFCC avec leurs dérivées première et deuxième en partant du principe que les delta et delta delta MFCC permettent de capturer les dynamiques temporelles du signal donc une meilleure compréhension de celui-ci, sachant que les paramètres acoustqiues sont normalisés et les hyperparamètres sont ajustés. Le tableau 4.9 ci-dessous présente les résultats obtenus.

Tableau 4. 9. Performances des noyaux SVM après combinaison des 13 paramètres MFCC et leur dérivée première et deuxième.

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	88.46%	79.81%	95.19%	91.35%
F1 score	88.31%	77.84%	95.07%	91.11%
Hyperparamètres	C: 10, degree: 2	'C': 1, 'gamma': 'scale'	C=10, gamma: 0.01	C=0.1

D'après les taux de reconnaissance fournis par chaque noyau, nous remarquons une diminution des performances lorsque on a combiné les 13 MFCC avec leurs dérivées première et deuxième, et l'ordre de performance des quatre noyaux reste toujours le même. Les deux figures 4.12 et 4.13 montrent respectivement une comparaison des noyaux SVM à l'issue de cette combinaison, et une comparaison des deux solutions MFCC seuls et MFCC avec leurs dérivées première et deuxième.

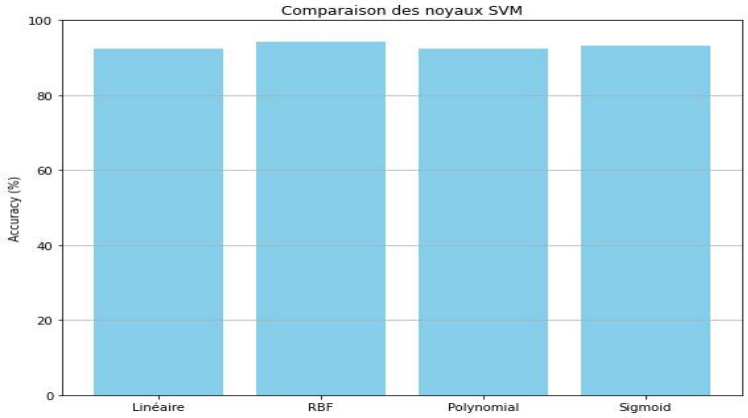

Figure 4. 12. Performances des noyaux SVM pour 13 MFCC et leur dérivée première et deuxième

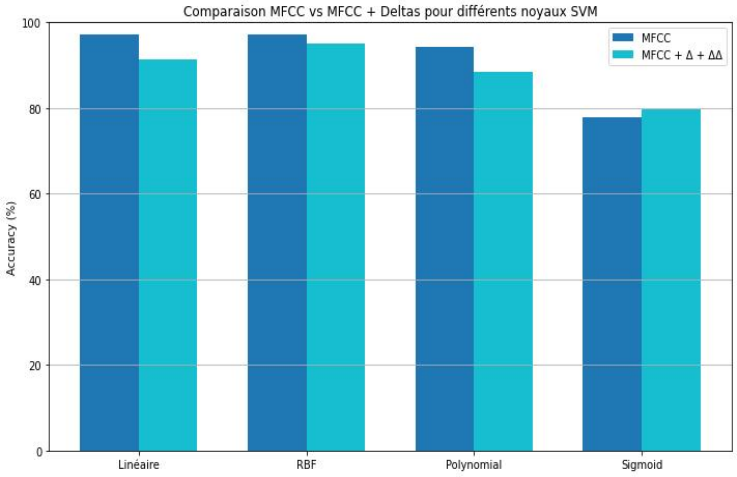

Figure 4. 13. Comparaison des performances des différents noyaux SVM pour MFCC seuls et MFCC combinés avec leur dérivées première et deuxième.

g) 40 MFCC et delta et delta delta MFCC

Le même test que le précédent a été effectué, mais en prenant 40 MFCC au lieu de 13 combinés avec les dérivées premières et deuxième. Le tableau 4.10 montre les résultats obtenus.

Tableau 4. 10. Performances des noyaux SVM après combinaison des 40 MFCC et leur dérivée première et deuxième

	Noyau Polynomial	Noyau Sigmoïde	Noyau RBF (Gaussien)	Linéaire
Précision	92.31%	93.27%	94.23%	92.31%
F1 score	92.31%	93.25%	94.27%	92.26%
Hyperparamètres	C: 10, degree: 2	'C': 1, 'gamma': 'auto'	C=10, gamma: scale	C=1

Par rapport aux résultats fournis par l'SVM pour différents noyaux, lorsque 40 MFCC seuls sont utilisés à son entrée, l'ajout de la première et deuxième dérivée de l'MFCC a fait baissé les performances, comme le montre le graphe présenté dans la figure 4.14 ci-dessous. Cependant,

pour le noyau polynomial et linéaire les mêmes performances sont obtenues. Par conséquent, cette combinaison n'est pas bénéfique, mais au contraire, elle ajoute de la complexité au système vu le nombre élevé de caractéristiques utilisé en entrée du classifieur (40MFCC, 40deltaMFCC, 40DeltaDeltaMFCC).

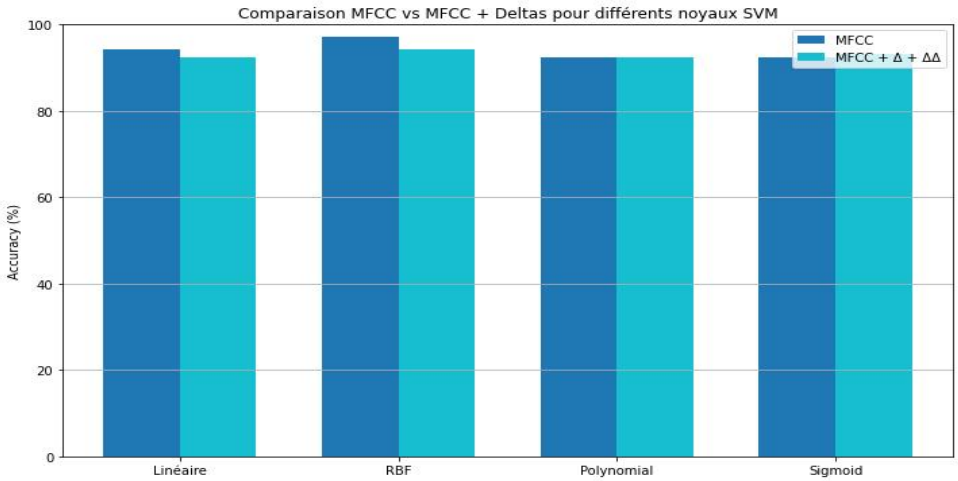

Figure 4. 14. Comparaison des performances des deux solutions : 40 MFCC seuls et MFCC +Δ+ΔΔ

Cette solution offre aussi des performances inférieures à celles obtenues en utilisant uniquement 13 MFCC et leurs dérivées première et deuxième et ceci pour le noyau RBF. Pour le reste des noyaux, cette solution est plus performante.

- **Discussion et interprétation**

Plusieurs et différents résultats peuvent être tirés de cette expérimentation, alors comment peut-on les expliquer ?

- Les résultats obtenus dans le cas de 3 classes sont très encourageants (85.19% pour le noyau polynomial, le noyau sigmoïde 59.26%, le noyau gaussien 96.30% et le noyau linéaire 100%) ; l'SVM linéaire présente les meilleurs résultats vient ensuite le noyau gaussien. Ces résultats sont les mêmes soit en utilisant 13 MFCC ou 40. Nous avons utilisé des valeurs par défaut des hyperparamètres pour les différents noyaux et sans normalisation des caractéristiques. Théoriquement, le choix entre un noyau linéaire et non linéaire est justifié par la nature des données à traiter ; c.à.d. pour des données non linéairement séparables l'utilisation d'un noyau non linéaire est nécessaire, tandis que pour des données séparables le noyau linéaire est suffisant et constitue le bon choix car il est moins coûteux en terme de calcul et de stockage vu qu'aucune transformation des données à un espace de plus grande dimension est faite. En effet, un taux de précision de 100% pour un SVM à noyau linéaire veut dire que nos descripteurs sont bien discriminants vu le nombre limité de classes. En revanche, pour les autres noyaux les taux obtenus sont expliqués par le choix arbitraire des hyperparamètres spécifiques à chacun de ces noyaux comme le paramètre C, gamma, et degré, mais aussi à la non normalisation des données. En réalité, le paramètre C qui joue un rôle important dans les

résultats à obtenir, et comme expliqué déjà dans le premier chapitre (chapitre 1), C est un paramètre de régularisation à fixer par l'utilisateur ! par conséquent, la valeur attribuée à C va influencer la taille de la marge et les erreurs de classification. En d'autres termes, une grande valeur de C conduit à une petite marge et inversement, lorsque la valeur de C est petite la marge sera grande.

- Dans le cas où le nombre de classes est de 7, mêmes résultats que pour le cas de 3 classes sauf que les taux de reconnaissances se dégradent (72.12% pour le noyau polynomial, 36.54% pour le noyau sigmoïde, 74.04% pour le noyau gaussien, 95.19% pour le linéaire), et l'ordre de classement de ces noyaux est toujours le même ; l'SVM linéaire puis vient le noyau gaussien, le noyau polynomial et enfin, le noyau sigmoïde. Ceci est expliqué par l'augmentation du nombre de classes, donc le problème devient plus complexe, en plus, le nombre des échantillons dans chaque classe est très petit. Par ailleurs, l'utilisation des données non normalisés et sans ajustement des hyperparamètres des noyaux va sûrement altérer négativement les performances du système.
- Lors de l'application de la validation croisée pour l'ajustement des hyperparamètres avec normalisation des descripteurs MFCC, nous avons pu atteindre les meilleures performances pour le noyau gaussien ou RBF (précision 97.12%t et f1 Score 97.11%) et ceci en prenant 40 MFCC en entrée de notre classifieur SVM. De même, lorsque nous prenons 13 MFCC en entrée avec normalisation, nous avons atteint les meilleures performances pour les deux noyaux RBF et Linéaire (une précision de 97.12% et 97.12% respectivement). Il convient de noter également que la normalisation des caractéristiques a augmenté considérablement les performances du noyau sigmoïde (de 36.54% à 92.31%), ceci est justifié par la forte sensibilité du noyau sigmoïde aux données non normalisées comme indiqué dans la littérature.
- Un dernier test consiste en l'ajout aux MFCC leur dérivée première et deuxième, dans le but d'augmenter les taux de reconnaissance. Cependant, cette combinaison a réduit les performances que ce soit en prenant uniquement les 13 MFCC avec leur dérivée première et deuxième ou les 40 MFCC avec aussi leur dérivée première et deuxième. Comme indiqué précédemment, nous avons opté pour l'ajout de la première et deuxième dérivée d'MFCC en partant de l'hypothèse que leur utilisation permet de capturer les dynamiques temporelles du signal donc une meilleure compréhension de celui-ci. En revanche, cette hypothèse s'applique à des sons supposés stationnaires, et comme la plupart des sons des classes à reconnaître sont impulsifs, l'ajout de la première et deuxième dérivée des paramètres MFCC va réduire les performances en ajoutant du bruit d'un côté, et rendre le système plus lent en augmentant le nombre de caractéristiques à traiter de l'autre côté.
- En plus de ces résultats, nous concluons aussi que pas seulement le nombre de classes qui a impact sur le taux de reconnaissance mais aussi la nature des classes à reconnaitre. Par exemple, dans le cas de 7 classes on n'obtient pas les mêmes résultats si on remplace une classe par une autre (la classe « fermeture de porte » par la classe « claquement de

porte » par exemple) ceci dépend de la ressemblance ou de la non ressemblance de la nouvelle classe avec les classes existantes, par conséquent, le taux de reconnaissance peut augmenter ou diminuer.

- En conclusion, dans cette expérimentation nous constatons que le noyau Gaussien ou RBF a permis de fournir les meilleurs résultats et son concurrent le noyau linéaire avec une différence de performance négligeable. Les MFCC montrent aussi leur efficacité pour la représentation pertinente des sons de la vie courante, mais la normalisation des données est une étape primordiale. Finalement, Un point important que nous pouvons constater est que le nombre des échantillons par classe ainsi que le nombre de caractéristiques influencent les performances de notre SVM. Selon des études précédentes et des travaux antérieurs, le noyau gaussien peut être utilisé lorsque le nombre des échantillons d'apprentissage est très élevé et le nombre de caractéristiques est petit. Cependant, le noyau linéaire offre de meilleurs résultats lorsque le nombre des exemples d'apprentissage est petit qui est notre cas dans cette expérimentation.

4.4. Conclusion

Dans ce chapitre, nous venons de présenter deux points importants dans un système de reconnaissance des sons. Le premier point, est une discussion des solutions possibles d'une architecture générale d'un système de reconnaissance de sons, ainsi que la proposition et l'adoption d'une architecture que nous voyons la plus appropriée par rapport aux solutions existantes. Le deuxième point est le sous-système de reconnaissance des sons dédié spécialement pour la reconnaissance des sons de l'environnement autres que la parole. Dans cette deuxième partie, nous avons également décrit la méthode d'extraction de caractéristiques qui sont les MFCC et la méthode de classification utilisé qui s'agit d'un classifieur SVM et ceci après une large étude comparative des travaux existants tels qu'ils sont décrits dans le chapitre 2.

La justification de chaque technique ou méthode utilisée dans ce travail est un point primordial de notre point de vue, et c'est la raison pour laquelle nous voyons important de mettre des justifications pour chaque pas.

Cette partie a été achevée par une expérimentation que je considère petite par rapport à nos objectifs, malgré les performances élevées que nous avons obtenues, mais qui constitue pour nous un point de départ pour des travaux plus importants. Une partie de cette expérimentation est publiée dans [Abdoune et al., 2024].

En fait, cette expérimentation décrit un système de classification des sons en utilisant les SVM pour la classification et les MFCC comme paramètres acoustiques. Dans la méthode de classification SVM nous avons comparé les résultats pour différents noyaux notamment, le noyau gaussien, le noyau polynomial, le noyau sigmoïde et le noyau ou l'SVM linéaire. La base de données utilisée dans ce travail est une base qui a été déjà utilisée dans des travaux antérieurs mais qui n'est pas assez volumineuse telle que décrite dans ce chapitre. Après avoir définir notre

corpus de sons dédié pour l'habitat dans le chapitre 3, nous avons essayé dans ce travail de cibler d'abord les sons de détresse qui sont les cris et les sons de bris de verre. Ensuite, nous avons étendu le test pour couvrir d'autres types de sons de notre corpus afin de voir le comportement du système et de discuter les résultats, et par conséquent ouvrir la voie pour de nouveaux tests et de nouvelles études et perspectives. Plusieurs conclusions peuvent être tirées de ce travail, que nous résumons ainsi :

- Pour le cas de 3 classes de sons notamment les cris, les bris de verre et les sons de vaisselle, nos tests sont fait en prenant des valeurs par défaut des hyperparamètres, donc sans ajustement, et aussi en prenant des vecteurs acoustiques bruts sans aucune normalisation des données, les résultats obtenus sont les suivants :
    - Le taux de reconnaissance obtenu pour l'SVM linéaire est de 100%, vient juste après le noyau RBF de 96.30%, puis le noyau polynomial de 85.19% et enfin, le noyau sigmoïde de 59.26%. L'SVM linéaire présente les meilleurs résultats vient ensuite le noyau gaussien. Ces résultats sont les mêmes, soit en utilisant 13 MFCC ou 40.
    - En général, avoir un taux de reconnaissance qui est défini ici à 100% est très satisfaisant. De plus, adopter un noyau linéaire constitue un bon choix en terme de puissance de calcul et de consommation de la mémoire.
- Pour le cas de 7 classes nous avons procédé ainsi :
    - Tests sans normalisation ni ajustement des hyperparamètres : l'ordre des noyaux selon le taux de reconnaissance est le même que celui des résultats obtenus avec 3 classes, mais avec une dégradation dans les taux de reconnaissance : 95.19% pour l'SVM linéaire, 74.04%pour le noyau gaussien, 72.12% pour le noyau polynomial, et 36.54% pour le noyau sigmoïde.
    - Tests avec normalisation des données et ajustement des hyperparamètres des différents noyaux par validation croisée : nous avons pu atteindre les meilleures performances par le noyau gaussien ou RBF avec une précision de 97.12% et ceci en prenant 40 MFCC en entrée de notre classifieur SVM. De même, pour 13 MFCC nous avons atteint les meilleures performances pour les deux noyaux RBF et linéaire d'une précision de 97.12%. Il convient de noter également que La normalisation des caractéristiques a augmenté considérablement les performances du noyau sigmoïde de 36.54% à 92.31%, ceci est justifié par la forte sensibilité du noyau sigmoïde aux données non normalisées comme indiqué dans la littérature. Donc, la normalisation des caractéristiques ainsi que l'ajustement des hyperparamètres des noyaux SVM sont essentiels pour obtenir une précision élevée.

- Le dernier test consiste en la combinaison des MFCC et leur dérivée première et deuxième, en utilisant d'abord 13 MFCC, puis 40, dans le but d'augmenter les taux de reconnaissance des classes de sons précédentes. Cependant, cette combinaison a réduit les performances. Comme indiqué précédemment, nous avons opté pour l'ajout de la

première et deuxième dérivée d'MFCC en partant de l'hypothèse que leur utilisation permet de capturer les dynamiques temporelles du signal donc une meilleure compréhension de celui-ci. En revanche, cette hypothèse s'applique à des sons supposés stationnaires, et comme la plupart des sons des classes à reconnaître sont impulsifs, l'ajout de la première et deuxième dérivée des paramètres MFCC n'est pas bénéfique et va réduire les performances en ajoutant du bruit d'un côté, et rendre le système plus lent en augmentant le nombre de caractéristiques à traiter de l'autre côté.

En conclusion, dans cette expérimentation nous constatons que le noyau Gaussien ou RBF a permis de fournir les meilleurs résultats et son concurrent le noyau linéaire avec une différence de performance négligeable. Les MFCC montrent aussi leur efficacité pour la représentation pertinente des sons de la vie courante.

Faire recours à une base de données testée déjà par d'autres auteurs au lieu d'utiliser une base de données personnelle est dans le but de comparaison avec d'autres systèmes de reconnaissance qui visent les mêmes classes de son. Ce point constitue pour nous un point important dans la démarche de ce travail. Cependant la comparaison de notre travail avec d'autres n'a pas eu lieu dans cette thèse pour de multiples raisons et que nous laissons en perspectives.

En effet, nous devons mentionner ici qu'en raison du petit nombre d'échantillons dans les classes de sons et de la variation du nombre d'échantillons d'une classe à l'autre, il serait préférable d'utiliser des techniques d'augmentation des données et d'équilibrage des classes pour obtenir des résultats plus précis. D'une part, l'augmentation des données sera employée pour accroître la diversité et le volume de l'ensemble de données en générant des variations supplémentaires des données existantes, ce qui contribuera à améliorer la robustesse et la généralisation du modèle. D'autre part, des techniques d'équilibrage des classes seront appliquées pour corriger tout déséquilibre existant entre les différentes classes de sons, en veillant à ce que le modèle soit formé sur une distribution plus équitable des classes. Ces améliorations devraient permettre d'obtenir de meilleures performances et des résultats plus fiables.

Il est important aussi de signaler que l'expérimentation présentée ici n'est qu'une partie validée du travail entier et plusieurs autres tests peuvent être effectués et dans différents niveaux du système entier,

- L'utilisation des méthodes de sélection des paramètres afin d'améliorer les résultats.
- Tester le système avec d'autres paramètres acoustiques et même le tester en combinant plusieurs autres paramètres autres que les MFCC et leur dérivée première et deuxième.
- Essayer d'autres méthodes de classifications telles que les réseaux de neurones et le deep learning afin de faire des comparaisons avec les SVM.
- Utiliser une base de données standard et plus riche afin de mieux tester le système et par conséquent obtenir des résultats plus réels.

En fin, l'enrichissement du système de télésurveillance par d'autres capteurs tels que les contacteurs de portes, l'infrarouge, l'accéléromètre et la collecte de ces données en combinaison avec le système de reconnaissance de sons peut nous conduire à un système complet qui permet la détection des activités des habitants et leur meilleure surveillance. Ce dernier point constitue le point de départ de cette thèse et reste comme perspective avec même une possibilité d'exploiter d'autres solutions dans la conception du système entier de télésurveillance.

Conclusion générale et perspectives

Les sons qu'on peut rencontrer dans l'environnement sont très variés et de différentes natures notamment, la parole, la musique et les sons environnementaux, sons extérieurs ou sons intérieurs, on parle alors du paysage sonore qui est considéré d'une grande richesse. Pour pouvoir reconnaitre ces différentes catégories de sons, des domaines distincts existent selon la nature du son à reconnaitre tels que la reconnaissance de la parole, la reconnaissance des sons de l'environnement et la reconnaissance du locuteur. En effet, la reconnaissance des sons peut être considérée comme le domaine le moins étudié par rapport à la reconnaissance de la parole et la reconnaissance du locuteur, et il est moins mature. Cependant, dans ces dernières années il est devenu plus actif et beaucoup de travaux ont été réalisés.

Le travail de recherche effectué dans cette thèse se situe dans le cadre général de la reconnaissance des sons de l'environnement, dans le but de recherche des meilleures méthodes de classification et d'extraction de caractéristiques pour la reconnaissance de sons, et de voir le comportement du système avec les méthodes usuelles appliquées à la reconnaissance de la parole, du locuteur et de la musique.

En effet, l'objectif du travail envisagé dès le départ est la conception d'un système de télésurveillance des personnes âgées en se basant sur le canal audio pour la détection des situations de détresse des habitants. Par conséquent, notre travail est une intersection de deux domaines, notamment les maisons intelligentes et en particulier la télésurveillance, et la reconnaissance des sons qui peut être considérée comme un domaine à part, mais aussi comme un outil pour atteindre et réaliser les objectifs tracés par la maison intelligente. Comme dans tout sujet de recherche, nous commençons par un objectif principal, et pour l'atteindre, plusieurs sous objectifs doivent être accomplis, et c'est notre cas dans cette thèse.

Notre **premier sous-objectif** est la proposition d'un corpus de sons de la vie courante à l'intérieur de la maison, puis collecte et construction d'une base de données. Cette partie a été décrite dans un chapitre à part, c'est le chapitre 3. Une petite expérimentation sur cette base de données a été réalisée afin de la tester. Ce travail représente notre première contribution et qui se trouve dans [Abdoune et Fezari, 2016].

Le **deuxième sous objectif** est l'analyse et l'étude des concepts fondamentaux de la théorie de la reconnaissance des sons, c'est ce que nous avons présenté dans le premier chapitre, il s'agit d'un dictionnaire pour les mots clés du domaine auquel les lecteurs peuvent se référer, en partant de la définition des sons jusqu'aux méthodes d'extraction des caractéristiques et de classification pour un système de reconnaissance des sons.

Un **3ème objectif**, souligné d'emblée, consiste à mettre en avant l'état de l'art sur la reconnaissance de sons en essayant de répondre aux questions : quelles méthodes ? Comment faire nos choix ? Est-ce que la reconnaissance des sons a ses propres méthodes ou bien elle est basée sur des méthodes utilisées dans le domaine de la reconnaissance de la parole et du locuteur. Une étude comparative a été menée, et il en ressort que la reconnaissance des sons se base sur les méthodes de classification utilisées pour la reconnaissance de la parole, et parmi les

méthodes de classification les plus prometteuses les machines à vecteurs support qui présentent les meilleurs résultats en comparaison avec les GMM, les réseaux de neurones et les HMM. Cependant, la comparaison n'a pas touché et ne peux pas toucher tous les travaux, donc nos conclusions restent approximatives et dépendent des travaux comparés. Quant aux méthodes d'extraction de caractéristiques, divers paramètres ont été comparés tels que les MFCC, les LPC, etc. Les MFCC montrent des résultats comparables à ceux issus d'autres paramètres. De plus, la combinaison des MFCC avec le delta MFCC et l'énergie donne de bons résultats. En partant du principe d'amélioration des taux de reconnaissances en jouant sur les paramètres acoustiques à utiliser, des conclusions sont tirées d'un point de vue théorique et statistique, mais l'expérimentation a porté uniquement sur les MFCC seuls et MFCC combinés avec leur première et deuxième dérivée, et de ce fait, la combinaison des paramètres a été laissée en perspectives. Le contenu de de cette partie est bien détaillé dans *le chapitre 2* qui est d'une grande importance dans ce manuscrit vu qu'il est consacré pour l'analyse et l'étude des travaux qui sont en relation avec notre sujet de recherche. En effet, Trois études comparatives ont été abordées dans ce chapitre :

1) **La première étude** concerne une synthèse et une comparaison des systèmes de reconnaissance de sons environnementaux intérieurs et extérieurs, cette comparaison a mis l'accent sur les variantes suivantes : les méthodes de classification utilisées, les méthodes d'extraction de caractéristiques, les méthodes de sélection des caractéristiques, les bases de données utilisées et les taux de reconnaissance atteints.
2) **La deuxième étude** a porté sur l'analyse et la comparaison des travaux de détection des situations de détresse via la reconnaissance des sons anormaux tels que les cris, les coups de feu dans un objectif de surveillance. Chaque travail définit ses sons d'intérêt, autrement dit, les sons indiquant une situation de danger. Différentes bases de données ont été utilisées pour chaque travail, mais aussi les méthodes de classification varient d'une application à une autre et il n'y a pas une méthode bien précise pour ce type d'application. Les taux de reconnaissance obtenus varient d'une application à une un autre et d'une méthode à l'autre dans le même travail. Quoique l'évaluation de l'efficacité des méthodes de classification pour la détection des sons anormaux ou de détresse repose principalement sur le taux de reconnaissance ou la précision. Cependant, lorsqu'il s'agit de comparer différents travaux entre eux, cette tâche devient complexe en raison de plusieurs facteurs.
    - Tout d'abord, chaque travail peut utiliser sa propre base de données ou une base de données différente, ce qui rend difficile une comparaison directe des performances. La composition et la qualité des bases de données peuvent influencer considérablement les résultats obtenus par chaque méthode de classification.
    - En outre, les classes de sons à reconnaître peuvent varier d'une application à l'autre, bien que l'objectif global reste le même (la détection des sons anormaux ou de détresse). Cette variabilité peut introduire des biais dans les comparaisons et rendre difficile l'établissement d'une référence commune pour l'évaluation des méthodes.

- Enfin, le choix des méthodes de classification implique souvent un compromis entre la précision des résultats obtenus et le coût de calcul nécessaire pour les atteindre. Des méthodes plus complexes et sophistiquées peuvent obtenir de meilleures performances, mais elles peuvent également nécessiter des ressources computationnelles importantes, ce qui peut être un facteur limitant dans certains contextes.
- Ainsi, pour mener une comparaison juste et significative des travaux de détection des sons anormaux ou de détresse, il est essentiel de prendre en compte ces défis et de bien choisir les bases de données, les critères d'évaluation, et les contraintes computationnelles. Les évaluations basées sur des données standardisées et des métriques bien définies peuvent contribuer à une meilleure compréhension des performances des différentes méthodes de classification dans ce domaine.

3) **La troisième étude** est un aperçu plutôt qu'une étude des travaux de reconnaissance de sons basés sur les méthodes d'apprentissage profond ou plus communément du deep learning. Selon les travaux présentés dans cette section, nous voyons l'importance de ces méthodes dans l'amélioration des taux de reconnaissance par rapport aux méthodes classiques. Cependant, certains cas montrent que des méthodes classiques puissantes tels que les SVM donnent des résultats comparables, voire supérieures. Sachant qu'il existe différentes méthodes du deep learning, mais chaque méthode donne un résultat différent qui peut être meilleur par rapport aux résultats obtenus par les autres méthodes classiques ou bien l'inverse. Nous constatons conséquemment, l'intérêt d'introduire ces méthodes dans nos futurs tests et travaux.

Le quatrième objectif est de faire des tests basés sur nos résultats et conclusions obtenues de nos études précédentes afin de valider nos choix et même d'examiner les résultats et d'en tirer des conclusions. C'est l'objet du *chapitre 4*, où nous avons d'abord présenté les différentes solutions proposées par les diverses architectures des systèmes de reconnaissance des sons, ensuite nous les avons critiquées et comparées en terme de faisabilité et de fiabilité pour enfin fixer une architecture qui sert de base pour notre système de classification de sons. Par la suite, nous avons abordé la classification des sons en se limitant à quelques classes de sons. Le système de classification en question, est un classifieur SVM avec comme paramètres les MFCC, et l'apprentissage et le test sont effectués sur le corpus de sons que nous avons présenté dans le chapitre 3. Dans un premier temps, nous avons testé uniquement 3 classes qui présentent en effet les classes de sons indiquant une situation de détresse notamment, les cris et les bris de verres et une classe des sons normaux qui est la vaisselle, ensuite nous avons étendu le nombre de classes en 7 classes : bris de verre, ouverture de porte, claquement de porte, cris, vaisselle, toux et écoulement d'eau, sachant que les tests sont effectués en prenant des valeurs par défaut des hyperparamètres, donc sans ajustement, et aussi en prenant des vecteurs acoustiques bruts sans aucune normalisation des données. Dans le cas de 3 classes, Les taux de reconnaissances obtenus pour l'SVM linéaire et l'SVM à noyau RBF sont les plus élevées de 100% et 96.30% respectivement, vient ensuite le noyau polynomial de 85.19% et enfin, le noyau sigmoïde de 59.26%. Avoir un taux de reconnaissance qui est défini ici à 100% est très satisfaisant, mais le

nombre de classes n'est pas assez grand. Lorsque le nombre de classes augmente les taux peuvent se dégrader et c'est le cas lors de l'augmentation du nombre de classes en 7 classes au lieu de 3. Les résultats obtenus dans ce dernier cas sont 95.19% pour l'SVM linéaire, 74.04%pour le noyau gaussien, 72.12% pour le noyau polynomial, et 36.54% pour le noyau sigmoïde. Après une phase de normalisation des données et d'ajustement des hyperparamètres par validation croisée nous avons pu atteindre les meilleures performances par le noyau gaussien ou RBF avec une précision de 97.12% et ceci en prenant 40 MFCC en entrée de notre classifieur SVM. De même, pour 13 MFCC nous avons atteint les meilleures performances pour les deux noyaux RBF et linéaire d'une précision de 97.12%. Il convient de noter également que La normalisation des caractéristiques a augmenté considérablement les performances du noyau sigmoïde de 36.54% à 92.31%, et ceci est justifié par la forte sensibilité du noyau sigmoïde aux données non normalisées. Le dernier test consiste en la combinaison des MFCC et leur dérivée première et deuxième, en utilisant d'abord 13 MFCC, puis 40, dans le but d'augmenter les taux de reconnaissance des classes de sons précédentes. Cependant, les performances sont dégradées avec cette combinaison. En effet, la combinaison MFCC et leur dérivée première et deuxième améliore la reconnaissance des sons qui sont de nature stationnaires, et comme la plupart des sons des classes à reconnaître sont impulsifs, l'ajout de la première et deuxième dérivée des paramètres MFCC n'est pas bénéfique et va réduire les performances en ajoutant du bruit d'un côté, et rendre le système plus lent en augmentant le nombre de caractéristiques à traiter de l'autre côté. Par conséquent, dans cette expérimentation nous constatons que l'SVM à noyau Gaussien ou RBF a permis de fournir les meilleurs résultats et son concurrent l'SVM à noyau linéaire avec une différence de performance négligeable. Les MFCC montrent aussi leur efficacité pour la représentation pertinente des sons de la vie courante.

En ce qui concerne *le chapitrer 3*, qui présente une proposition d'un corpus de sons pour la vie courante, plusieurs points nécessitent d'être discutés et expliqués ici :

La première démarche de notre travail est la création d'un corpus de sons et une base de données pour l'évaluation du système final, étant donné qu'au début de ce travail ils n'existaient pas des bases de données standards pour le benchmarking des applications, ce qui nous a incité à entreprendre la création d'une base de donnée sur laquelle nous avons fait de simples tests sans étude préalable sur les paramètres acoustiques à utiliser ni le type du classifieur à adopter. Après une étude approfondie sur la reconnaissance des sons et avec l'avancement des travaux dans ce domaine, nous avons pensé à utiliser une base de données déjà employée dans des travaux antérieurs en vue d'une possible comparaison. Pour cette raison, la base de données [6] a été utilisée, mais qui a aussi une taille limitée, néanmoins des résultats encourageants ont été obtenus.

En effet, les tests effectués dans cette expérimentation ne sont pas complets et ne couvrent qu'une partie de nos objectifs soulignés, mais considérés comme une porte d'entrées pour des solutions et tests plus approfondis et en particulier avec des bases de données standards et

volumineuses. De ce fait, beaucoup de tests restent à achever dans de futurs travaux que nous pouvons résumer ainsi :

- Utiliser une base de données standard et plus riche afin de mieux tester le système et par conséquent obtenir des résultats plus réels.
- Augmenter le nombre de classes
- Tester le système avec d'autres paramètres acoustiques et même le tester en combinant plusieurs autres paramètres.
- L'utilisation des méthodes de sélection des paramètres afin d'améliorer les résultats.
- Essayer d'autres méthodes de classification telles que les méthodes basées sur le deep learning afin de faire des comparaisons avec les SVM, et même passer à la combinaison des classifieurs ou les classifieurs hybrides.

Notre étude dans cette thèse a plusieurs implications pratiques pour les applications du monde réel. Tout d'abord, l'analyse de la littérature fournit une vue d'ensemble des méthodologies existantes et des progrès réalisés dans la technologie de reconnaissance des sons, ce qui favorise la poursuite de la recherche et du développement et suggère des axes d'amélioration. En plus, le système de reconnaissance sonore peut être intégré dans diverses applications de surveillance, telles que les systèmes de sécurité, les environnements de soins de santé et les milieux industriels, cela dépend des sons à reconnaître : sons de détresse pour les soins de santé, alarmes dans les industries, etc. En outre, l'intégration du système dans des environnements intelligents tels que les maisons et les villes intelligentes en facilitant, par exemple, les systèmes d'alerte basés sur les sons détectés.

En plus des propositions faites dans ce manuscrit, ainsi que les perspectives citées précédemment, ce travail peut se poursuivre dans plusieurs directions à la fois théoriques et techniques. Citons ici des perspectives qui sont moins visibles mais qui sont d'une grande importance :

- Une **perspective** importante de cette thèse qui est considérée comme un des objectifs globaux de notre travail consiste en l'utilisation de différents capteurs dans le smart home afin d'atteindre des résultats plus exacts et ceci via la fusion de données multimodales venant des capteurs hétérogènes en utilisant des techniques de fusion des données. Via l'utilisation de différents capteurs et avec la fusion de données, le système final sera plus fiable en réduisant l'incertitude et en résolvant les conflits. Un travail similaire peut être trouvé sur [Medjahed, 2011].
- Un autre travail aussi important qui présente lui-même un sujet de recherché à part est le choix de l'architecture des applications de reconnaissance des sons. En effet, une grande partie des applications de reconnaissance vocale, de classification musicale et d'indexation audio sont prises en charge par les plates-formes PC, le cloud computing et / ou les téléphones intelligents puissants, par contre la plupart des applications pour la maison intelligente sont intégrées dans un produit matériel dont la puissance de calcul

ne peut pas correspondre à celle d'un PC. Il serait donc bénéfique et important de pouvoir comparer les solutions existantes et essayer de trouver un compromis entre le prix du matériel, la complexité des applications déployées sur ce matériel, les ressources nécessaires pour l'exécution et la qualité du service fournis. A notre avis, faire recours au cloud computing est une bonne solution mais il faut faire face aux problèmes liés à la confidentialité, la bande passante et la latence, etc.

- Faire une comparaison du système proposé avec des systèmes existants. En effet, Effectuer une comparaison du système proposé avec les systèmes existants se révèle être une tâche complexe dans le domaine de la reconnaissance de son, principalement en raison de plusieurs facteurs. Tout d'abord, le nombre de classes de sons auxquelles chaque système s'intéresse peut varier considérablement, certaines applications se concentrant sur une seule classe, tandis que d'autres en couvrent plusieurs. De plus, la nature des sons d'intérêt (impulsionnels, stationnaires, ou une combinaison des deux) peut également varier d'un système à l'autre. De même, les catégories de sons étudiées peuvent varier, ajoutant encore à la complexité de la comparaison. Enfin, le recours à des bases de données différentes pour chaque système rend difficile la comparaison entre les diverses approches réalisées, compliquant ainsi l'identification des méthodes les plus adaptées à la reconnaissance de son.
- En effet, les applications en temps réel telles que la surveillance audio et les systèmes de sécurité nécessitent une reconnaissance rapide et en temps réel. Obtenir des résultats précis en temps réel, en particulier avec des ressources informatiques limitées, est un défi. Cela nécessite souvent des algorithmes efficaces qui équilibrent la précision et la vitesse, mais ce compromis peut limiter la qualité de la reconnaissance. C'est pourquoi le choix de la méthode de classification et des caractéristiques audio est une tâche essentielle et c'est la raison pour laquelle nous avons entrepris cette recherche.
- Prendre en considération l'aspect « sujet âgé ». Il est primordial d'expliquer que dans cette thèse aucune considération n'a été faite sur la population visée par cette application. Autrement dit, parler de la surveillance des personnes âgées ou handicapées ou tout autre type de personne ne va rien changer dans notre système tant que la reconnaissance de la parole n'a pas fait objet de cette thèse. C'est dans la reconnaissance de la parole que notre système varie avec la variation des personnes visées. Le choix des paramètres acoustiques dépend fortement de la nature de la personne concernée où on doit tenir compte des caractéristiques spécifiques.
- **En fin**, La nature des sons environnementaux, leur diversité et leur structure qui n'est pas assez simple à comprendre vu l'absence de phonèmes et qui rends leur modélisation difficile, en plus de l'aspect impulsif, stationnaire ou non stationnaire, tout ça nous a fait penser à chercher d'autres solutions en dehors des paramètres utilisés pour la reconnaissance de la parole. Un point important qui nous a attiré l'attention est de sortir de la présentation par frame en suscitant un changement de mentalité concernant la définition des événements sonores. Plutôt que de considérer un son comme des ensembles de trames acoustiques instantanées et de s'appuyer uniquement sur des

techniques de classification trame par trame, le but est d'adopter une perspective plus large qui traite les événements sonores comme des séquences acoustiques interrompues. Cette approche vise à combiner la modélisation acoustique instantanée avec une prise en compte plus globale des aspects temporels et structuraux, permettant ainsi une intégration plus cohérente des deux éléments, comme il a été mentionné dans [Krstulović, 2018].

- L'amélioration de l'exactitude et de la précision en explorant des techniques avancées d'apprentissage automatique. De plus, l'optimisation de la solution pour un traitement plus rapide et en temps réel et la mise à l'échelle du système pour gérer des flux de données à haut débit.
- Tester d'autres représentations 2D telle que le spectrogramme qui capture à la fois les caractéristiques temporelles et fréquentielles. En outre, le recours à de nouvelles méthodes d'apprentissage profond nécessite de travailler avec des bases de données plus importantes, ce qui peut être fait en appliquant des techniques d'augmentation des données, en particulier pour les types de sons difficiles à obtenir tels que la chute de personnes. En conclusion, exploiter des méthodes matures dédiées pour la reconnaissance d'image telles que les CNN avec utilisation des spectrogrammes peut présenter une solution au problème de la diversité et la variation des sons de l'environnement et leur manque d'une structure claire.
- Intégrer le système de reconnaissance sonore à d'autres technologies, telles que les appareils IoT ou les applications pilotées par l'IA.

Pour conclure, L'obtention d'un taux de reconnaissance élevé pour un système de reconnaissance des sons en utilisant des méthodes de classification et d'extraction de caractéristiques quelconques ne signifie pas le bon choix de ces méthodes tant qu'il existe d'autres critères importants sur lesquels on peut se baser pour évaluer ces systèmes. Un de ces critères est le temps de calcul nécessaire pour l'exécution de ces méthodes ainsi que la taille des paramètres acoustiques utilisés ; savoir si son exécution nécessite un matériel avec des puissances élevées de traitement ou bien un simple matériel peut être utilisé est très important. De ce fait, cette thèse est le début d'un processus de recherche qui nécessite d'être complété dans de futurs travaux, mais aussi ouvre la voie à de multiples sujets de recherche. En continuant dans ce même sujet, notre premier objectif consiste en l'intégration de ce travail dans l'IoT (Internet of Things), en utilisant le smart Home comme un exemple et en mettant en œuvre une solution embarquée.

Références bibliographiques

Références bibliographiques

[Abdoune et al., 2024] Abdoune, L., Fezari, M., Dib, A. (2024). Indoor Sound Classification with Support Vector Machines: State of the Art and Experimentation. International Journal of Computational Methods and Experimental Measurements 12(3):269-279 .https://www.iieta.org/journals/ijcmem/paper/10.18280/ijcmem.120307 https://doi.org/10.18280/ijcmem.120307.

[Abdoune et Fezari, 2014] Abdoune, L., & Fezari, M. (2014, January). A sound database for health smart home. In *2014 World Congress on Computer Applications and Information Systems (WCCAIS)* (pp. 1-5). IEEE.

[Abdoune et Fezari, 2016] Abdoune, L., & Fezari, M. (2016). Everyday life sounds database: telemonitoring of elderly or disabled. *Journal of Intelligent Systems*, *25*(1), 71-84.

[Abdoune et Fezari, 2017] Abdoune, L., & Fezari, M. (2017). Feature extraction for everyday life sounds. 5 th International Conference on Control & Signal Processing (CSP-2017), Proceeding of Engineering and Technology –PET, Vol.26 pp.186-191.

[Agostini et al., 2001] Agostini, G., Longari, M., & Pollastri, E. (2003). Musical instrument timbres classification with spectral features, in IEEE Fourth Workshop on Multimedia Signal Processing, pp. 97–102.

[Agrawal et al., 2017] Agrawal, D. M., Sailor, H. B., Soni, M. H., & Patil, H. A. (2017, August). Novel TEO-based Gammatone features for environmental sound classification. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1809-1813). IEEE.

[Alías et al., 2016] Alías, F., Socoró, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, *6*(5), 143.

[Al-Karawi, 2019] Al-Karawi, K. A. (2019). Robustness speaker recognition based on feature space in clean and noisy condition. *International Journal of Sensors Wireless Communications and Control*, *9*(4), 497-506.

[AlQahtani et al., 2010] AlQahtani, M. O., Muhammad, G., & Alotaibi, Y. A. (2010, July). Environment sound recognition using zero crossing features and MPEG-7. In *2010 Fifth International Conference on Digital Information Management (ICDIM)* (pp. 502-506). IEEE.

[André-Obrecht, 1988] André-Obrecht R. (1988). A new statistical approach for automatic speech segmentation, *IEEE Transanctions on Acoustics, Speech and Signal Processing*, vol. 36, no. 1, pp. 29–40.

[Arslan, 2017] Arslan, Y. (2017). Impulsive Sound Detection by a Novel Energy Formula and its Usage for Gunshot Recognition. *arXiv preprint arXiv:1706.08759*.

[Azlan et al., 2005] Azlan M., Cartwright I., Jones N., Quirk T., & West G. (2005). Multimodal monitoring of the aged in their own homes. In *Proceedings of the International Conference on Smart Homes and Health Telematics (ICOST)*. IOS Press, 264–271.

[Bach et al., 2011] Bach, J.H.; Anemüller, J.; Kollmeier, B. (2011). Robust speech detection in real acoustic backgrounds with perceptually motivated features. Speech Communication. 53, 690–706.

[Baker et al., 2009] Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., & O'Shaughnessy, D. (2009). Developments and directions in speech recognition and understanding, part 1 [dsp education]. Signal Processing Magazine, IEEE, 26(3) :75–80.

[Boddapati et al., 2017] Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia computer science*, *112*, 2048-2056.

[Bolón-Canedo et al., 2013] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems*, *34*, 483-519.

[Bonhomme, 2008] Bonhomme, S. (2008). *Méthodologie et outils pour la conception d'un habitat intelligent* (Doctoral dissertation, Institut National Polytechnique de Toulouse-INPT).

[Bregman, 1990] Bregman A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press, Cambridge, Mass, USA.

[Büchler et al., 2005] Büchler, M., Allegro, S., Launer, S., & Dillier, N. (2005). Sound classification in hearing aids inspired by auditory scene analysis. *EURASIP Journal on Advances in Signal Processing*, *2005*(18), 387845.

[Burges, 1998] Burges C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 955-974.

[Cakir et al., 2017] Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., Virtanen, T., Cakir, E., ... & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *25*(6), 1291-1303.

[Castelli et al., 2003] Castelli E., Vacher M., Istrate D., Besacier L. & Sérignat J. F. (2003, july). Habitat telemonitoring system based on the sound surveillance, in: 1st International

Conference on Information Communication Technologies in Health, ISBN 960-813-17-1, pp. 141 – 146, Greece.

[Chachada et Kuo, 2013] Chachada, S., & Kuo, C. C. J. (2013). Environmental sound recognition: A survey. *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013*.

[Chachada et Kuo,2014] Chachada, S., & Kuo, C. C. J. (2014). Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, *3*, e14.

[Chan et al., 2003] Chan, M., Campo, E., & Estève, D. (2003, september). PROSAFE, a multisensory remote monitoring system for the elderly or the handicapped. *Proceedings of the ICOST*, *3*, 89-95.

[Chan et al., 2008] Chan, M., Estève, D., Escriba, C., & Campo, E. (2008). A review of smart homes—Present state and future challenges. *Computer methods and programs in biomedicine*, *91*(1), 55-81.

[Chechik et al., 2008] Chechik, G., Ie, E., Rehn, M., Bengio, S., & Lyon, D. (2008, October). Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (pp. 105-112).

[Chu et al., 2006] Chu, S., Narayanan, S., Kuo, C. C. J., & Mataric, M. J. (2006, July). Where am I? Scene recognition for mobile robots using audio features. In *2006 IEEE International conference on multimedia and expo* (pp. 885-888). IEEE.

[Chu et al., 2009] Chu, S., Narayanan, S., & Kuo, C. C. J. (2009). Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1142-1158.

[Chung et al., 2018] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*.

[Cortes et Vapnik, 1995] Cortes, C. & Vapnik, V. (1995). Support vector networks. Machine Learning, 20:273–297. http://dx.doi.org/10.1007/BF00994018

[Cowling et Sitte, 2002] Cowling, M. and Sitte, R. (2002). Recognition of environmental sounds using speech recognition techniques. In Wysocki, T., Darnell, M., and Honary, B., editors, Advanced Signal Processing for Communication Systems, volume 703 of The International Series in Engineering and Computer Science, pages 31–46. Springer US.

[Cowling et Sitte, 2003] Cowling, M. & Sitte, R. (2003). Comparison of techniques for environmental sound recognition, *Pattern Recognition Letters*, vol.24, pp.2895-2907.

[Cowling, 2004] Cowling, M. (2004). Non-Speech Environmental Sound Classification System for Autonomous Surveillance. Thesis (PhD Doctorate), Griffith University, Brisbane.

[Crocco et al., 2016] Crocco, M., Cristani, M., Trucco, A., & Murino, V. (2016). Audio surveillance: A systematic review. *ACM Computing Surveys (CSUR)*, *48*(4), 52.

[Dash et Liu, 2003] Dash M, Liu H (2003) Consistency-based search in feature selection. J Artif Intell 151(1–2):155–176.

[Davis et Mermelstein, 1980] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366.

[Delgado-Contreras et al., 2014a] Delgado-Contreras, J.R., Garcia-Vazquez, J.P. & Brena, R.F. (2014). Classification of environmental audio signals using statistical time and frequency features. In: Electronics, Communications and Computers (CONIELECOMP), 2014 International Conference, pp. 212–216 http://dx.doi.org/10.1109/CONIELECOMP. 2014.6808593.

[Delgado-Contreras et al., 2014b] Delgado-Contreras, J.R., Garcia-Vazquez, J.P., Brena, R.F., Galván-Tejada, C.E., Galván-Tejada, J.I. (2014). *Feature selection for place classification through environmental sounds*. Procedia Computer Science 37, 40-47 .

[Dennis et al., 2011] Dennis, J., Tran, H. D., and Li, H. (2011). Spectrogram image feature for sound event classification in mismatched conditions. Signal Processing Letters, IEEE, 18(2) :130–133.

[Dennis et al., 2013a] Dennis, J., Tran, H., and Chng, E. (2013a). Overlapping sound event recognition using local spectrogram features and the generalised hough transform. Pattern Recognition Letters, 34(9) :1085 – 1093.

[Dennis et al., 2013b] Dennis, J., Tran, H. D., and Chng, E.-S. (2013b). Image feature representation of the subband power distribution for robust sound event classification. Audio, Speech, and Language Processing, IEEE Transactions on, 21(2) :367–377.

[DiCarlo et Cove, 2009] DiCarlo, A.S., MSW Glen Cove (2009). Smart Homes -Home Automation- (Livable New York Resource Manual), New York Research Center, New York City, USA.13.Ch.

[Dietterich et Bakiri, 1995] Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, *2*, 263-286, arXiv preprint cs/9501101.

[Dufaux et al., 2000] Dufaux, A., Besacier, L., Ansorge, M., & Pellandini, F. (2000, September). Automatic sound detection and recognition for noisy environment. In *2000 10th European Signal Processing Confere nce* (pp. 1-4). IEEE.

[Dufaux, 2001] Dufaux, A. (2001). *Detection and recognition of impulsive sound signals* (Doctoral dissertation, Verlag nicht ermittelbar).

[Essid, 2005] Essid, S. (2005). *Classification automatique des signaux audio-fréquences: reconnaissance des instruments de musique* (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).

[Favory et al., 2018] Favory, X., Fonseca, E., Font, F., & Serra, X. (2018, November). Facilitating the Manual Annotation of Sounds When Using Large Taxonomies. In *Proceedings of the 23rd Conference of Open Innovations Association FRUCT* (p. 60). FRUCT Oy.

[Fleury et al., 2010] Fleury, A., Vacher, M., Portet, F., Chahuara, P., & Noury, N. (2010, May). A Multimodal Corpus Recorded in a Health Smart Home. Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, LREC, 18-21 May 2010, Malta, pp. 99-105.

[Foggia et al., 2016] Foggia, P., Petkov, N., Saggese, A., Strisciuglio, N., & Vento, M. (2016). Audio surveillance of roads: A system for detecting anomalous sounds. *IEEE transactions on intelligent transportation systems*, *17*(1), 279-288.

[Font, 2013] Font, F., Roma, G., & Serra, X. (2013). Freesound technical demo. In Proceedings of the ACM International Conference on Multimedia, pages 411–412. ACM.

[Fukunaga, 1990] Fukunaga, K. (1990). Feature Extraction and Linear Mapping for Classification. *Introduction to Statistical Pattern Recognition.* Academic Press, Inc. California, USA.

[Gaver, 1993] Gaver, W. W. (1993). What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, *5*(1), 1-29.

[Gemmeke et al., 2017] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 776-780). IEEE.

[Gerhard, 2003] Gerhard, D. (2003). Audio signal classification: History and current techniques. Department of Computer Science, University of Regina, Regina, Saskatchewan, CANADA.

[Goldhor, 1993] Goldhor, R. S. (1993, April). Recognition of environmental sounds. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 149-152). IEEE.

[Gouda et al., 2018] Gouda, S. K., Kanetkar, S., Harrison, D., & Warmuth, M. K. (2018). Speech recognition: keyword spotting through image recognition. *arXiv preprint arXiv:1803.03759*.

[Greco et al., 2020] Greco, A., Petkov, N., Saggese, A., & Vento, M. (2020). AReN: a deep learning approach for sound event recognition using a brain inspired representation. *IEEE Transactions on Information Forensics and Security*, *15*, 3610-3624.

[Gururani et al., 2018] Gururani, S., Summers, C., & Lerch, A. (2018). Instrument activity detection in polyphonic music using deep neural networks. In *International Society for Music Information Retrieval (ISMIR)*.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. J Mach Learn 46(1–3):389–422.

[Hall, 1990] Hall, M., A. (1999). Correlation-based feature selection for machine learning. PhD thesis, University of Waikato, Hamilton.

[Han et al., 2017] Han, Y., Kim, J., Lee, K., Han, Y., Kim, J., & Lee, K. (2017). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *25*(1), 208-221.

[Hang et al., 2019] Hang, T., Feng, J., Li, X., & Yan, L. (2019, January). Water Sound Recognition Based on Support Vector Machine. In *International Conference on Ubiquitous Information Management and Communication* (pp. 986-995). Springer, Cham.

[Harper, 2003] R. Harper (2003). Inside the smart home, pp. 17. Springer, London.

[Hsu et al., 2003] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

[Hsu et al., 2010] C.W Hsu, C.-C. Chang, C.-J. Lin, (2010).LIBSVM : a library for support vector machines. *Technical Report*, Department of Computer Science and Engineering, National Taiwan University, Taiwan.

[Hsu et Lin, 2002] Hsu, C.-W., & Lin., C.-J. (2002).A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks, 13, 415-425.

[Huang et al., 2001] Huang, X., Acero, A., Hon, H. W., & Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

[Ibarz et al., 2008] Ibarz, A., Bauer, G., Casas, R., Marco, A., & Lukowicz, P. (2008, October). Design and evaluation of a sound based water flow measurement system. In *European Conference on Smart Sensing and Context* (pp. 41-54). Springer, Berlin, Heidelberg.

[Istrate et al., 2004] Istrate, D., Vacher, M., Castelli, E., & Nguyen, C. P. (2004, September). Sound processing for health smart home. In *Proc. International Conference on Smart homes and health ITelematics*, Singapour, 15-17 (pp. 41-48).

[Istrate et al., 2006] Istrate, D., Castelli, E., Vacher,M., Besacier, L., and Serignat, J.-F. (2006). Information extraction fromsound for medical telemonitoring. *Information Technology in Biomedicine, IEEE Transactions on*, 10(2) :264–274.

[Istrate et al., 2008] Istrate, D., Vacher, M., Serignat, J.F. (2008). Embedded implementation of distress situation identification through sound analysis. The Journal on Information Technology in Healthcare 6, 204–211.

[Istrate, 2003] Istrate, D. (2003). *Détection et reconnaissance des sons pour la surveillance médicale* (Doctoral dissertation, Institut National Polytechnique de Grenoble-INPG).

[Ito et al., 2009] Ito, A., Aiba, A., Ito, M., & Makino, S. (2009, August). Detection of abnormal sound using multi-stage GMM for surveillance microphone. In *2009 Fifth International Conference on Information Assurance and Security* (Vol. 1, pp. 733-736). IEEE.

[Jesudhas et Ranjan, 2024] Jesudhas, P. P., & Ranjan, P. V. (2024). A novel approach to build a low complexity smart sound recognition system for domestic environment. Applied Acoustics, 221, 110028.

[Jović et al., 2015] Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 1200-1205). Ieee.

[Kefauver, 1999] Kefauver, A. (1999). The Digital Encoding Process. *Fundamentals of Digital Audio. (Volume 14: The computer music and digital audio series)*. A-R Editions, Inc. Madison, Wisconsin, U.S.A.

[Kim et al., 2020] Kim, J., Min, K., Jung, M., & Chi, S. (2020). Occupant behavior monitoring and emergency event detection in single-person households using deep learning-based sound recognition. *Building and Environment*, *181*, 107092.

[Kinnunen et Li, 2011] Kinnunen, T., Li, H. (2011). An overview of text-independent speaker recognition: From features to supervectors. Speech Communication. 52 (1), 12–40.

[Kour et Mehan, 2015] Kour, G., & Mehan, N. (2015). Music genre classification using MFCC, SVM and BPNN. *International Journal of Computer Applications*, *112*(6).

[Krstulović, 2018] Krstulović, S. (2018). Audio event recognition in the smart home. *Computational Analysis of Sound Scenes and Events*, 335-371. https://doi.org/10.1007/978-3-319-63450-0_12

[kuklyte et al., 2009] Kuklyte, J., Kelly, P., Ó Conaire, C., O'Connor, N. E., & Xu, L. Q. (2009). Anti-social behavior detection in audio-visual surveillance systems, *The Workshop on Pattern Recognition and Artifcial Intelligence for Human Behavior Analysis*, Reggio Emilia, Italy.

[Lecomte et al., 2011] Lecomte, S., Lengellé, R., Richard, C., Capman, F., & Ravera, B. (2011, August). Abnormal events detection using unsupervised one-class svm-application to audio surveillance and evaluation. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 124-129). IEEE.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature 521 (7553), 436-444.

[Lee et al., 2021] Lee, Y., Lim, S., Kwak, I.Y. (2021). CNN-based acoustic scene classification system. Electronics, 10(4): 371. https://doi.org/10.3390/electronics10040371

[Lu et al., 2006] Lu, L., Liu, D., Zhang, H.J. (2006). Automatic Mood Detection and Tracking of Music Audio Signals. IEEE Trans. Audio Speech Lang. Process. 14, 5–18.

[Lyon, 2011] Lyon, R.F. (2011). A Survey of Audio-Based Music Classification and Annotation. IEEE Trans. Multimedia, 13, 303–319.

[Ma, et al., 2002] Ma, J., Zhao, Y., & Ahalt, S. (2002). OSU SVM Classifier Matlab Toolbox, Ohio State University, USA. 2002.

[Maldonado et al., 2009] Maldonado, S., & Weber, R. (2009). A wrapper method for feature selection using support vector machines. *Information Sciences*, *179*(13), 2208-2217.

[Marković et al., 2017] Marković, B., Galić, J., Grozdić, Đ., Jovičić, S. T., & Mijić, M. (2017). Whispered speech recognition based on gammatone filterbank cepstral coefficients. *Journal of Communications Technology and Electronics*, *62*(11), 1255-1261.

[McLoughlin et al., 2017] McLoughlin, I., Zhang, H., Xie, Z., Song, Y., Xiao, W., & Phan, H. (2017). Continuous robust sound event classification using time-frequency features and deep learning. *PloS one*, *12*(9), e0182309.

[Medjahed, 2011] Medjahed, H., Istrate, D., Boudy, J., Baldinger, J. L., & Dorizzi, B. (2011, June). A pervasive multi-sensor data fusion for smart home healthcare monitoring. In *2011 IEEE international conference on fuzzy systems (FUZZ-IEEE 2011)* (pp. 1466-1473). IEEE.

[Mejía-Lavalle et al., 2006] Mejía-Lavalle M, Sucar E, Arroyo G (2006). Feature selection with a perceptron neural net. In: Proceedings of the international workshop on feature selection for data mining, pp 131–135

[Min et al., 2019] Min, K., Jung, M., Kim, J., & Chi, S. (2019). Sound Event Recognition-Based Classification Model for Automated Emergency Detection in Indoor Environment. In *Advances in Informatics and Computing in Civil and Construction Engineering* (pp. 529-535). Springer, Cham.

[Mitrović et al., 2010] Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Features for content-based audio retrieval. In *Advances in computers* (Vol. 78, pp. 71-150). Elsevier.

[Moncrieff et al., 2007] Moncrieff, S., Venkatesh, S., & West, G. (2007). Online audio background determination for complex audio environments. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *3*(2), 8-es.

[Muhammad et al., 2010] Muhammad, G., Alotaibi, Y. A., Alsulaiman, M., & Huda, M. N. (2010, June). Environment recognition using selected MPEG-7 audio features and mel-frequency cepstral coefficients. In *2010 Fifth international conference on digital telecommunications* (pp. 11-16). IEEE.

[Muhammad et Alghathbar, 2009] Muhammad, G., & Alghathbar, K. (2009, December). Environment recognition from audio using MPEG-7 features. In *2009 Fourth International Conference on Embedded and Multimedia Computing* (pp. 1-6). IEEE.

[Murthy et al., 1999] Murthy, H. A., Beaufays, F., Heck, L. P., & Weintraub, M. (1999). Robust text-independent speaker identification over telephone channels. *IEEE Transactions on Speech and Audio Processing*, *7*(5), 554-568.

[Nanni et al., 2017] Nanni, L., Costa, Y. M. G., Lucio, D. R., Silla, C. N., & Brahnam, S. (2017). *Combining visual and acoustic features for audio classification tasks. Pattern Recognition Letters, 88, 49–56.* doi:10.1016/j.patrec.2017.01.013

[Ntalampiras et al., 2008] Ntalampiras, S., Potamitis, I., & Fakotakis, N. (2009, April). On acoustic surveillance of hazardous situations. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 165-168). IEEE.

[Park et Yoo, 2020] Park, H., & Yoo, C. D. (2020). CNN-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification. *IEEE Signal Processing Letters*, *27*, 411-415.

[Patterson et al., 1987] Patterson, R., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). An efficient auditory filterbank based on the gammatone function. In a meeting of the IOC Speech Group on Auditory Modelling at RSRE, volume 2.

[Patterson et al., 1995] Patterson, R. D., Allerhand, M. H., and Giguère, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. The Journal of the Acoustical Society of America, 98(4) :1890–1894.

[Peng et al., 2005] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, *27*(8), 1226-1238.

[Piczak, 2015] Piczak, K. J. (2015, October). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015-1018).

[Pieraccini, 2012] Pieraccini, R. (2012). The Voice in the Machine. Building Computers That Understand Speech; MIT Press: Cambridge, MA, USA.

[Plumbley et al., 2018] Virtanen, T., Plumbley, M. D., & Ellis, D. (Eds.). (2018). *Computational analysis of sound scenes and events*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-63450-0

[Qi et al., 2013] Qi, J., Wang, D., Jiang, Y., & Liu, R. (2013, May). Auditory features based on gammatone filters for robust speech recognition. In *2013 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 305-308). IEEE.

[Rabaoui et al., 2007] Rabaoui, A., Davy, M., Rossignol, S., Lachiri, Z., Ellouze, N., & équipe SequeL, I. F. (2007, September). Sélection de descripteurs audio pour la classification des sons environnementaux avec des SVM mono-classe. In *Actes du 21eme Colloque GRETSI: Traitement du Signal et des Images (GRETSI'07)*.

[Rabaoui et al., 2008] Rabaoui, A., Davy, M., Rossignol, S., & Ellouze, N. (2008). Using one-class SVM and wavelets for audio surveillance. *IEEE Transactions on information forensics and security*, *3*(4), 763-775.

[Radhakrishnan et al., 2005] Radhakrishnan, R., Divakaran, A., & Smaragdis, A. (2005, October). Audio analysis for surveillance applications. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* (pp. 158-161). IEEE.

[Ratanpara et Patel, 2015] Ratanpara, T., & Patel, N. (2015). Singer identification using MFCC and LPC coefficients from Indian video songs. In *Emerging ICT for Bridging the Future-*

Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1 (pp. 275-282). Springer, Cham.

[Ren et al., 2015] Ren, J. M., Wu, M. J., & Jang, J. S. R. (2015). Automatic music mood classification based on timbre and modulation features. *IEEE Transactions on Affective Computing*, *6*(3), 236-246.

[Rosenthal et Okuno, 1998] Rosenthal, D. F., & Okuno, H. G. (1998). *Computational auditory scene analysis*. Lawrence Erlbaum Associates Publishers.

[Rouas et al., 2006] Rouas, J. L., Louradour, J., & Ambellouis, S. (2006, September). Audio events detection in public transport vehicle. In *2006 IEEE Intelligent Transportation Systems Conference* (pp. 733-738). IEEE.

[Russo et al., 2019] Russo, M., Stella, M., Sikora, M., & Pekić, V. (2019). Robust cochlear-model-based speech recognition. *Computers*, *8*(1), 5.

[Sarno et al., 2018] Sarno, R., Ridoean, J. A., Sunaryono, D., & Wijaya, D. R. (2018). Classification of Music Mood Using MPEG-7 Audio Features and SVM with Confidence Interval. *International Journal on Artificial Intelligence Tools*, *27*(05), 1850016.

[Schafer, 1993] Schafer, R. M. (1993). *The soundscape: Our sonic environment and the tuning of the world*. Simon and Schuster.

[Schalkoff, 1990] Schalkoff, R.J. (1990). *Pattern Recognition: Statistical, Structural and Neural Approaches.* John Wiley & Sons, Inc. New York, NY, U.S.A.

[Sehili, 2013] Sehili, M. E. A. (2013). *Reconnaissance des sons de l'environnement dans un contexte domotique* (Doctoral dissertation, Institut National des Télécommunications).

[Sharan et Moir, 2016] Sharan, R. V., & Moir, T. J. (2016). An overview of applications and advancements in automatic sound recognition. *Neurocomputing*, *200*, 22-34.

[Shie et Chen, 1999] Qian, S., & Chen, D. (1999). Joint time-frequency analysis. *IEEE signal processing magazine*, *16*(2), 52-67. New York, NY, U.S.A.

[Sigtia et al., 2016] Sigtia, S., Stark, A. M., Krstulović, S., & Plumbley, M. D. (2016). Automatic environmental sound recognition: Performance versus computational cost. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2096-2107.

[Smeaton et McHugh, 2005] Smeaton, A. F., & McHugh, M. (2005, November). Towards event detection in an audio-based sensor network. In *Proceedings of the third ACM international workshop on Video surveillance & sensor networks* (pp. 87-94).

[Stolcke et al., 2007] Stolcke, A., Kajarekar, S. S., Ferrer, L., & Shrinberg, E. (2007). Speaker recognition with session variability normalization based on MLLR adaptation transforms. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 1987-1998.

[Sueur, 2018] Sueur, J. (2018). A very short introduction to sound analysis for those who like elephant trumpet calls or other wildlife sound.

[Taherian et al., 2020] Taherian, H., Wang, Z. Q., Chang, J., & Wang, D. (2020). Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1293-1302.

[Tipler, 1991] Tipler, P. (1991). Sound, *Physics for scientists and engineers*. Third Edition; Extended Edition. Worth Publishers Inc, New york, USA. 1991, Chap.14; p439-p483.

[Uzkent et al., 2012] Uzkent, B., Barkana, B. D., & Cevikalp, H. (2012). Non-speech environmental sound classification using SVM with a new set of features. *International Journal of Innovative Computing, Information and Control*, *8*(5), 3511-3524.

[Vacher et al., 2003] Vacher, M., Istrate, D., Besacier, L., Serignat, J. F., & Castelli, E. (2003, July). Life sounds extraction and classification in noisy environment. In *5th IASTED-SIP*.

[Vacher et al., 2004] Vacher, M., Istrate, D., and Serignat, J. (2004). Sound detection and classification through transient models using wavelet coefficient trees. In LTD, S., editor, *Proc. 12th European Signal Processing Conference*, pages 1171–1174, Vienna, Austria.

[Vacher et al., 2010a] Vacher, M., Fleury, A., Portet, F., Serignat, J. F., & Noury, N. (2010). Complete sound and speech recognition system for health smart homes: application to the recognition of activities of daily living. *New Developments in Biomedical Engineering*, (pp. 645-676).

[Vacher et al., 2010b] Vacher, M., Portet, F., Fleury, A., & Noury, N. (2010, July). Challenges in the processing of audio channels for ambient assisted living. In *The 12th IEEE International Conference on e-Health Networking, Applications and Services* (pp. 330-337). IEEE.

[Vacher et al., 2011] Vacher, M., Portet, F., Fleury, A., & Noury, N. (2011). Development of audio sensing technology for ambient assisted living : Applications and challenges. International Journal of E-Health and Medical Communications, 2(1) :35–54. 2011.

[Vacher, 2011] Vacher, M. (2011). *Analyse sonore et multimodale dans le domaine de l'assistance à domicile* (Doctoral dissertation, Université de Grenoble).

[Valenzise et al., 2007] Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., & Sarti, A. (2007, September). Scream and gunshot detection and localization for audio-surveillance systems.

In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on* (pp. 21-26). IEEE.

[Valero et Alías, 2012] Valero, X. & Alías, F. (2012). Gammatone wavelet features for sound classification in surveillance applications. In Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, pages 1658–1662. IEEE.

[Vaufreydaz et al., 2000] Vaufreydaz, D., Bergamini, C., Serignat, J.-F., Besacier, L. & Akbar, M. (2000). A new methodology for speech corpora definition from internet documents, *LREC'2000, 2nd Int. Conference on Language Ressources and Evaluation*, Athens, Greece, pp. 423–426.

[Vert, 2001] Vert, J.P. (2001). Introduction to Support Vector Machines and Applications to Computational Biology. *Seminar Report*, Kyoto University, Japan.

[Wang et al., 2008] Wang, J. C., Lee, H. P., Wang, J. F., & Lin, C. B. (2008). Robust environmental sound recognition for home automation. *IEEE transactions on automation science and engineering*, *5*(1), 25-31.

[Watanabe et al., 2018] Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825-13835.

[Witten et Frank, 2005] Witten, I.H. & Frank, E. (2005). Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Francisco.

[Xiong et al., 2018] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018, April). The Microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5934-5938). IEEE.

[Yamakawa et al., 2011] Yamakawa, N., Takahashi, T., Kitahara, T., Ogata, T., & Okuno, H. G. (2011). Environmental sound recognition for robot audition using matching-pursuit. In *Modern Approaches in Applied Intelligence: 24th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2011, Syracuse, NY, USA, June 28–July 1, 2011, Proceedings, Part II 24* (pp. 1-10). Springer Berlin Heidelberg.

[You et Li, 2012] You, G., & Li, Y. (2012, October). Environmental sounds recognition using tespar. In *2012 5th International Congress on Image and Signal Processing* (pp. 1796-1800). IEEE.

[Zhang et al., 2005] Zhang, D., Gatica-Perez, D., Bengio, S., & McCowan, I. (2005, June). Semi-supervised adapted HMM for unusual event detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 611-618). IEEE.

[Zhang et al., 2016] Zhang, W., Lei, W., Xu, X., & Xing, X. (2016). Improved Music Genre Classification with Convolutional Neural Networks. In *INTERSPEECH* (pp. 3304-3308).

Webographie

[1]: Le projet SWEET-HOME : Système Domotique d'Assistance au Domicile, SWEET-HOME : système domotique d'assistance au domicile - Archive ouverte HAL

[2]: « *PROSAFE project: Système complet de surveillance de personnes âgées* » – RNTS -, http:\\www.laas.fr\PROSAFE.

[3]: Home care for Dependent Elderly People - Educational Path for Informal Carers, http://homecareproject.eu/, 2019.

[4]: impulsive sound, https://encyclopedia2.thefreedictionary.com/impulsive+sound, 2019.

[5]: Understanding FFTs and Windowing, Understanding FFTs and Windowing - NI, 2024.

[6] : noise of life, https://github.com/amsehili/noise-of-life , 2017

[7]: "BBC Sound Effects Library—Original Series", https://www.sound-ideas.com/Product/152/BBC-Sound-Effects-Library-Original-Series

CONTRIBUTIONS SCIENTIFIQUES

1. Publications dans des journaux internationaux

1. Abdoune, L., & Fezari, M. (2016). Everyday life sounds database: telemonitoring of elderly or disabled. Journal of Intelligent Systems, 25(1), 71-84. https://www.degruyter.com/document/doi/10.1515/jisys-2014-0110/html

   Intitulé de la Revue : Journal of Intelligent Systems (Jisys)
   ISSN : …0334-1860 EISSN : ……2191-026X
   Indexation de la revue : Scopus , Web of Science, ..
   Editeur : WALTER DE GRUYTER GMBH , GENTHINER STRASSE 13, BERLIN, GERMANY, D-1078.
   H index: 31
   Impact factor: 2.1

2. Abdoune, L., et al. (2024). Indoor Sound Classification with Support Vector Machines: State of the Art and Experimentation. International Journal of Computational Methods and Experimental Measurements (IJCMEM), Vol. 12, No. 3, September, 2024, pp. 269-279, Journal homepage: http://iieta.org/journals/ijcmem

   Intitulé de la Revue : International Journal of Computational Methods and Experimental Measurements (IJCMEM)
   ISSN : …2046-0546 EISSN : 2046-0554
   Editeur : Giulio Lorenzin, International Information and Engineering Technology Association
   Indexation de la revue : Scopus, SCImago (SJR), DOAJ, CrossRef, Portico, EBSCOhost, ProQuest, Cabell's Directory, British Library, Library of Congress, Google Scholar, Dimensions, CAB Abstracts (CABI), Zetoc, MIAR, Microsoft Academic, CNKI Scholar, Baidu Scholar.
   H index : 12

2. Communications avec comité de lecture international

1. Abdoune, L., & FEZARI, M. (2011). Detection and Classification of Abnormal Sounds in a Habitat. *Procedings of STA*, Decembre, 2011
2. Abdoune, L., & Fezari, M. (2014, January). A sound database for health smart home. In *2014 World Congress on Computer Applications and Information Systems (WCCAIS)* (pp. 1-5). IEEE.
3. Abdoune, L., & Fezari, M. (2017). Feature extraction for everyday life sounds. 5 th International Conference on Control & Signal Processing (CSP-2017), Proceeding of Engineering and Technology –PET, Vol.26 pp.186-191.