Analyse sonore et multimodale dans le domaine de l'assistance à domicile

Michel Vacher

UNIVERSITÉ DE GRENOBLE

MEMOIRE

Pour obtenir le diplôme d'

HABILITATION A DIRIGER DES RECHERCHES

Spécialité : **Informatique et Mathématiques Appliquées**

Arrêté ministériel : 25 avril 2002

Présenté par

Michel VACHER

Travaux préparés au sein du **Laboratoire d'Informatique de Grenoble**

Analyse sonore et multimodale dans le domaine de l'assistance à domicile

Thèse soutenue publiquement le **18 octobre 2011**,
devant le jury composé de :

Mme Michèle ROMBAUT
Professeur des Universités, Université Joseph Fourier, Grenoble I, Président
M. Jean-Paul HATON
Professeur des Universités, Institut Universitaire de France, Nancy, Rapporteur
M. Corneliu BURILEANU
Professeur, Université POLITEHNICA, Bucarest, Roumanie, Rapporteur
Mme Frédérique LAFOREST
Professeur des Universités, Université Jean Monnet, Télécom St Etienne, Rapporteur
M. Jean CAELEN
Directeur de Recherche CNRS, LIG, Grenoble, Examinateur
M. Christian BOITET
Professeur des Universités, Université Joseph Fourier, Grenoble I, Examinateur
M. Alain FRANCO
Professeur des Universités, CHU de Nice, Examinateur

Au cours de ma carrière de chercheur, j'ai eu l'occasion de travailler avec de nombreux collègues ; par ailleurs, une partie des activités décrites dans ce manuscrit s'est faite à travers l'encadrement d'étudiants en IUT, Master, thèse CNAM ou Doctorat. Qu'ils en soient tous chaleureusement remerciés même si je n'ai pas pu tous les citer.

J'aimerai également remercier les professeurs Jean-Paul Haton, Corneliu Burileanu et Frédérique Laforest pour m'avoir fait l'honneur de rapporter ce mémoire. J'exprime aussi mes vifs remerciements envers les professeurs Christian Boitet, Alain Franco et Michèle Rombaut et le docteur Jean Caelen pour avoir accepté d'être examinateurs de cette HDR.

Je remercie Christian Boitet, Laurent Besacier et Hervé Blanchon qui m'ont appuyé dans cette initiative et ont eu la gentillesse de relire ce mémoire.

Je remercie tout particulièrement Jean-François Serignat, ancien responsable de l'équipe GEOD, qui m'a accueilli dans son équipe et formé à la reconnaissance automatique de la parole avant de me confier la responsabilité du thème Habitat Intelligent pour la Santé. De la même manière, je remercie Laurent Besacier, Jean Caelen, Eric Castelli, Denis Tuffelli, Dominique Vaufreydaz, anciens membres de l'équipe GEOD, pour leur accueil et leurs conseils avisés.

Je remercie Vincent Rialle du laboratoire TIMC et Norbert Noury de l'INL avec qui j'ai eu pendant plusieurs années une collaboration fructueuse et enrichissante sur l'Habitat Intelligent pour la Santé.

Je remercie François Portet, Solange Rossato, Brigitte Meillon, Jean-Claude Durand, Christian Perrot, Francis Jambon, Bernard Cassagne avec qui j'ai toujours le plaisir de collaborer dans des projets ou d'avoir des discusions techniques et scientifiques fructueuses. Je fais de même pour Benjamin Lecouteux qui vient juste de nous rejoindre.

Je remercie les étudiants avec qui j'ai pu travailler lors de la préparation de leur thèse ou mémoire CNAM, et avec qui j'ai beaucoup appris : Dan Istrate, Anthony Fleury, Hubert Glasson, Pedro Chahuara, Frédéric Aman, Stéphane Chaillol et Sylvain Méniard.

Je remercie Nicolas Gac, Noë Guirand et Remus Dugheanu pour la qualité du travail accompli lors de leur Master, travail qui a permis à chaque fois de défricher un domaine nouveau pour l'équipe.

Je remercie Anthony Fleury, Pedro Chahuara et Benjamin Lecouteux, pour leur implication importante qui a permis des avancées scientifiques significatives dans cette thématique de recherche, et avec qui j'ai beaucoup appris.

J'exprime aussi ma gratitude au professeur Robert Goutte qui m'avait suivi et encouragé lors de la préparation de ma thèse, il y a déjà de cela de nombreuses années, et m'a ainsi transmis sa passion de la recherche.

Table des matières

Abstract

The average age of the population in industrialized countries is steadily increasing. Seniors living alone are more numerous, either because they prefer to live independently, or because of a lack of space in the specialized institutions. We must find solutions allowing them to continue to stay at home comfortably and safely. Smart housings can constitute one of these solutions.

One of the biggest challenges in Ambient Assisted Living (AAL) is to develop smart homes that anticipate the health needs of their inhabitants while maintaining their safety and comfort. It is therefore essential to facilitate interaction with the smart home through systems that respond naturally to voice commands, using microphones but no tactile interfaces.

This thesis defines the concept of smart home and describes some interesting projects. It then explains how home assistance can take advantage of this concept thanks to sound analysis.

The acceptability of a voice interface as part of the intelligent home has been studied through an experiment that showed what are the wishes, expectations and fears of older users, their families, and social workers.

Audio analysis in smart homes is a still unexplored research area, the interest and the methods to analyze sound information in a smart home have been studied here in an experiment which helped to highlight the challenges and technological obstacles to be removed in order to use sound information in addition to other modalities and, in the case of speech, distant speech recognition (ASR in remote recording conditions).

A practical solution using several microphones is then presented. The intended purpose is to achieve a voice control system which will allow the user to control their environment not only by conventional switches and remote control devices, but also by voice.

The advantage of the audio information combined with that of home automation sensors is then revealed through a multimodal analysis making it possible to locate a person in a smart home or to determine their activity. Localization is necessary, for example to determine the context in which a home automation command has been issued. The activity can be used to observe changes in the habits of a person to assist in diagnosis.

Finally, the thesis presents the perspectives of research and future projects of the author. It is accompanied by the reproduction of four scientific papers published in refereed selective conferences.

Résumé

La moyenne d'âge de la population des pays industriels augmente régulièrement. Les personnes âgées vivant seules sont de plus en plus nombreuses, soit parce qu'elles préfèrent vivre de manière autonome, soit par manque de place dans les institutions spécialisées. Il faut donc trouver des solutions leur permettant de continuer à rester chez elles de manière confortable et sûre. Les habitats intelligents peuvent constituer une de ces solutions.

Un des plus grands défis dans l'Assistance à la Vie Autonome (AVA) est de concevoir des habitats intelligents pour la santé qui anticipent les besoins de leurs habitants tout en maintenant leur sécurité et leur confort. Il est donc essentiel de faciliter l'interaction avec l'habitat intelligent grâce à des systèmes qui réagissent naturellement aux commandes vocales, en utilisant des microphones et pas des interfaces tactiles.

Ce mémoire définit le concept de maison intelligente et présente quelques projets intéressants. Il précise ensuite de quelle manière l'assistance à domicile peut tirer parti de ce concept en s'appuyant sur l'analyse sonore.

L'acceptabilité d'une interface vocale dans le cadre de l'habitat intelligent a été étudiée grâce à une expérience qui a montré quels étaient les souhaits, les attentes et les craintes des utilisateurs âgés, de leurs familles, et des travailleurs sociaux.

L'analyse audio dans la maison intelligente étant un domaine de recherche encore peu exploré, l'intérêt et la manière d'analyser les informations sonores dans un habitat intelligent sont ensuite abordés par une expérience qui a permis de mettre en évidence les défis et les verrous technologiques qui devront être levés pour pouvoir utiliser les informations sonores en complément des autres modalités, et, dans le cas de la parole, la reconnaissance en conditions d'enregistrement distant.

Une solution pratique mettant en œuvre plusieurs microphones est ensuite présentée. Le but envisagé est la réalisation d'un système de commande vocale mettant l'utilisateur en mesure de piloter son environnement non seulement par les interrupteurs et télécommandes classiques, mais aussi par la voix.

L'intérêt de l'information audio combinée à celle des capteurs domotiques est ensuite mis en évidence au travers d'une analyse multimodale permettant de localiser une personne dans un habitat intelligent ou de déterminer son activité. La localisation est nécessaire, par exemple pour avoir connaissance du contexte dans lequel un ordre domotique a été donné. L'activité peut être utilisée pour observer une évolution des habitudes de la personne pour aider à un diagnostic.

Pour finir, le mémoire présente les perspectives de recherche et les projets à venir de l'auteur. Il est accompagné de la reproduction de 4 communications scientifiques publiées dans des congrès sélectifs à comité de lecture.

Table des figures

Liste des tableaux

Préambule

Contexte de l'assistance à domicile

La moyenne d'âge de la population des pays industriels augmente régulièrement du fait de l'amélioration des conditions de vie et des progrès de la médecine. L'Organisation des Nations Unies prévoit qu'en 2050 22% de la population dépassera 65 ans dans ces pays. Chaque état devra donc se préparer pour faire face à cette évolution et permettre à chacun de vivre dans les meilleures conditions possibles.

Différents projets de recherche étudient des solutions qui devraient permettre aux personnes âgées de continuer à rester chez elles le plus longtemps possible. Actuellement, les familles sont de plus en plus dispersées géographiquement du fait des emplois occupés par les enfants. Les personnes âgées vivent donc souvent isolées de leur famille et doivent rester autonomes. Par ailleurs, étant donné l'augmentation continuelle de l'espérance de vie, des troubles comme la maladie d'Alzheimer deviennent de plus en plus fréquents. Les solutions de télé-assistance les plus connues du grand public ont comme objectif de détecter des situations de détresse (en général la chute car c'est une crainte très fortement exprimée par les personnes âgées) ou d'assister la personne dans la prise régulière des médicaments prescrits.

Il faut aussi prendre en compte le fait que les personnes âgées sont certes fragiles mais ne sont pas toutes atteintes de pathologies. L'assistance aux personnes doit donc se situer dans le contexte plus général des maisons intelligentes qui visent à aider les personnes à garder une bonne maîtrise de leur environnement.

Composition du mémoire

Mes travaux, bien que fortement orientés vers l'analyse audio, qui concerne le parole et les sons de la vie courante, sont par nature pluridisciplinaires comme cela apparaîtra clairement au cours des différents chapitres de ce mémoire. Il convient de noter que l'analyse audio dans la maison intelligente est un domaine de recherche encore peu exploré. Les travaux présentés représentent la part la plus significative et la plus récente de mon activité des dix dernières années. Ce mémoire a été rédigé sur une période allant de janvier 2011 à mai 2011.

Les travaux et le projet de recherche présenté ont trait au domaine de l'assistance à domicile. Il m'a donc paru important de bien définir le concept de maison intelligente et de son rapport à l'assistance à domicile, c'est l'objet du chapitre 1. La façon dont l'assistance à domicile s'appuie sur la maison intelligente est abordée au cours de ce chapitre qui présente aussi quelques projets qui m'ont paru particulièrement intéressants. J'ai d'ailleurs participé

à certains d'entre eux, les projets RESIDE-HIS et DESDHIS en ce qui concerne l'analyse sonore.

Le chapitre 2 aborde l'intérêt et la manière d'analyser les informations sonores dans une maison intelligente. Une expérience permet de mettre en évidence les défis et les verrous technologiques qui devront être levés pour pouvoir utiliser les informations sonores en complément des autres modalités, et, dans le cas de la parole, la reconnaissance en conditions distantes.

L'assistance à domicile ne peut se concevoir sans que l'on ne prenne en compte les souhaits, les attentes et les craintes des utilisateurs. Le chapitre 3 présente donc une étude qui a été menée concernant l'acceptabilité d'une interface vocale dans le cadre du bâtiment intelligent.

Le chapitre 4 fait écho au chapitre 2, et, présente quelques pistes concernant la reconnaissance de la parole en conditions distantes grâce à l'utilisation de plusieurs microphones répartis dans un appartement. Le but envisagé est la réalisation d'un système de commande vocale mettant l'utilisateur en mesure de piloter son environnement, non seulement par les interrupteurs et télécommandes classiques, mais aussi par la voix.

Ensuite, les chapitres 5 et 6 sont consacrés à l'analyse multimodale, et, présenteront chacun une méthode qui a été mise en œuvre, l'une afin de localiser une personne dans un habitat intelligent, et l'autre pour déterminer son activité. La localisation est nécessaire, par exemple pour avoir connaissance du contexte dans lequel un ordre domotique a été donné. L'activité peut être utilisée pour observer une évolution des habitudes de la personne, de façon à aider à un diagnostic.

Pour finir, je présente mes perspectives de recherche et mes projets à venir. Quatre communications scientifiques publiées dans des congrès sélectifs à comité de lecture sont reproduites en annexe.

CHAPITRE 1

Domaine de l'assistance à domicile

1.1 Concept de maison intelligente

Une *maison intelligente* ou *smart home* (Aldrich, 2003; Chan et al., 2008) est une résidence équipée de technologie d'informatique ambiante (Weiser, 1991), qui anticipe et répond aux besoins de ses occupants en essayant de gérer de manière optimale leur confort et leur sécurité par action sur la maison, et en mettant en œuvre des connexions avec le monde extérieur. La maison intelligente est donc une branche importante de la *domotique*.

Le bâtiment intelligent est souvent au centre des projets européens lié à l'Assistance à la Vie Autonome (AVA), ou *Ambient Assisted Living (AAL)*, étant donné qu'il répond à certains besoins d'assistance médicale (Haigh and Yanco, 2002; Lacombe et al., 2005; Patterson et al., 2002; Tang and Venables, 2000; Rialle, 2007b). En effet, l'assistance aux personnes en perte d'autonomie est un objectif prioritaire de la commission européenne (6e et 7e PCRD). Parmi ces projets internationaux, on peut citer RoboCare (Bahadori et al., 2004), CompanionAble (Badii and Boudy, 2009) et COGNIRON [1] (Li and Wrede, 2007) qui utilisent un robot assistant communiquant avec la maison intelligente, ALADIN (Maier and Kempter, 2010) qui étudie l'effet de la lumière sur le bien être, ou encore EMERGE [2] qui vise à détecter quand les personnes s'écartent de leurs habitudes.

On peut diviser ces projets en plusieurs classes : ceux qui se concentrent sur la sécurité (chute essentiellement) tels que CAALYX (Anastasiou et al., 2008), ENABLE (Panek et al., 2007) ou CARE [3], ceux qui visent le développement de solutions logicielles dédiées à l'aide quotidienne aux personnes âgées tels que INHOME Vergados et al. (2008), NETCARITY [4] (focalisé sur l'inclusion sociale), HERMES (Jianmin Jiang and Zhang, 2009) (pour l'aide mémoire), PERSONA (Soler et al., 2010; Amoretti et al., 2010) ou encore SHARE-IT (Cortes et al., 2010), et ceux qui s'intéressent aux questions éthiques ou aux solutions globales tels que SENIOR [5] ou CAPSIL [6].

De nombreux laboratoires de recherche (publics ou privés) comme Domus (Giroux et al., 2002), le CEA/LETI et Orange Labs travaillent sur l'aide au maintien des personnes à domi-

1. http://www.cogniron.org/final/Home.php
2. http://www.emerge-project.eu/
3. http://care-aal.eu/
4. http://www.netcarity.org/
5. http://seniorproject.eu/
6. http://www.capsil.org/

cile. Les systèmes envisagés sont souvent très complexes (présence de nombreux capteurs), adaptés à une pathologie et un besoin particulier, ou bien impliquent que le patient porte en permanence un capteur fixé sur le corps, par exemple un capteur de chute ou un accéléromètre. Peu présente sur la scène internationale, la recherche française est assez active sur la scène nationale. On peut citer le projet GER'HOME [7] (Zouba et al., 2009), associant notamment l'INRIA et le CSTB dont le but était de détecter les activités des personnes en fonction de différents capteurs dont de la vidéo, ou encore PROSAFE (Bonhomme et al., 2008) ou CASPER [8] pour le suivi des personnes âgées.

1.2 Objectifs généraux de la maison intelligente

Si l'objectif de maintien de la santé et de compensation de la dépendance représente une très forte motivation des maisons intelligentes, il n'en demeure pas moins que leur développement nécessite des apports majeurs du monde industriel de la domotique (ex. : capteurs, réseaux, normes de communication, etc.) en dehors de toutes considérations médicales. C'est pourquoi les applications du bâtiment intelligent ont fait l'objet de projets menés d'une part par les industriels impliqués dans la production des appareils électriques et électroniques, comme Hager, Legrand, Philips, et Schneider Electric, et d'autre part par des constructeurs de composants électroniques comme AnalogDevice et Freescale, ainsi que par des fournisseurs d'énergie tels que EDF et d'autres PMI (Lutolf, 1992) dès les années 1990. Que ce soit via des projets de recherche européens, industriels ou de laboratoires, le nombre de solutions proposées reste très important, avec, parmi beaucoup d'autres : DomoNet [9], oBIX [10] (pour normaliser l'échange d'information entre smart homes), OpenRemote [11], EnTiMid (Nain et al., 2009), Attraco (Goumopoulos et al., 2008), Pobicos [12], etc.

Ibarz et al. (2007) se sont focalisés sur une partie spécifique de l'environnement de tous les jours, le grand électroménager (*white goods*) au travers du projet EASY LINE+ . Leur but était de développer une nouvelle génération d'appareils électroménagers ayant une utilisation plus facile et sans danger pour les personnes âgées ou handicapées, en partie en utilisant des RFID. Par ailleurs, les plate-formes d'expérimentation ont littéralement fleuri ces dernières années. Par exemple, l'équipe MULTICOM du LIG propose aux entreprises de les accompagner dans la conception et l'évaluation de services interactifs innovants dans l'environnement ubiquitaire baptisé DOMUS. On peut aussi citer l'initiative Open Living Lab [13] qui regroupe plus de 100 laboratoires d'expérimentation in vivo (Jambon, 2009), ou *living labs*, à travers le monde. Ces laboratoires d'expérimentation in vivo sont des plate-formes de développement de nouvelles technologies de la communication centrées sur l'utilisateur.

7. http://gerhome.cstb.fr/
8. http://www-prima.inrialpes.fr/casper/
9. http://sourceforge.net/projects/domonet/
10. http://www.obix.org/
11. http://www.openremote.org/
12. http://www.ict-pobicos.eu/
13. http://www.openlivinglabs.eu/

Un des centres d'expérimentation les plus connus est PlaceLab House_n [14], mis en place par le MIT et la société nord-américaine TIAX LCC, qui comporte plus de 400 capteurs de différentes natures (infrarouge, RFID, vidéo etc.) et qui est destiné à tester les nouvelles technologies. Il convient aussi de citer Adaptive house (Mozer, 2005) dont le but était de concevoir une maison intelligente pour observer les habitudes des occupants afin d'apprendre comment anticiper leurs besoins et y répondre. Le contrôle était effectué par une méthode combinant des réseaux de neurones et de l'apprentissage par renforcement. Dans le projet MavHome (Cook et al., 2003), le système avait des buts prédéfinis tels que minimiser le coût de maintenance de la maison et maximiser son confort. Le projet de maison intelligente UMASS (Lesser et al., 1999) s'attachait a développer une application multi-agent afin de contrôler les appareils électroménagers selon une perspective de consommation et de coordination.

1.3 Assistance à domicile au travers de la maison intelligente

L'assistance à domicile est devenue un enjeu important, étant donnée la part croissante de la population âgée, principalement dans les pays industrialisés. Cela correspond à plusieurs contraintes fortes : tout d'abord, l'augmentation forte du nombre de seniors et le souhait manifesté par ces personnes de vivre indépendamment le plus longtemps possible à leur domicile dans de bonnes conditions, ensuite le moindre coût pour la société lorsque les personnes peuvent continuer à vivre de manière autonome, ainsi que le manque de places disponibles dans des institutions spécialisées.

L'autonomie de la personne est couramment évaluée par les gériatres en utilisant une échelle décrivant l'aptitude ou l'inaptitude de la personne à s'assumer elle-même (Katz and Akpom, 1976) dans les Activités de la Vie Quotidienne (AVQ) ou *Activities of Daily Living* (ADL) qui sont l'ensemble des activités que chacun effectue naturellement dans la vie de tous les jours. Ces activités incluent le fait de se laver, de se nourrir, de se reposer...

Une autre échelle, ayant les mêmes champs d'application, prend en compte les instruments de la vie quotidienne et la capacité de la personne à les utiliser. Cette échelle se base sur les IADL (*Instrumented Activities of Daily Living*) définis par Lawton and Brody (1969). En France, la Grille AGGIR (Autonomie Gérontologie - Groupes Iso-Ressources) est utilisée pour évaluer l'autonomie physique et psychique grâce à l'observation des activités effectuées par la personne âgée, toute seule, ce qui permet de définir un groupe iso-ressource caractéristique du degré de dépendance.

Par ailleurs, Chan et al. (2009) ont passé en revue les projets ayant une perspective médicale et identifié les obstacles à surmonter, à savoir :

- les besoins des utilisateurs, l'acceptabilité et leur satisfaction ;
- la fiabilité et l'efficacité des systèmes de capteurs et des logiciels de traitement ;
- la standardisation de l'information et des systèmes de communication ;
- les contraintes légales et éthiques ;

14. http://architecture.mit.edu/house_n/placelab.html

– les réductions de coût et les impacts socioéconomiques.

Par ailleurs, plusieurs études ont été conduites pour définir les besoins des personnes âgées vis à vis d'un système pouvant les aider dans leur vie de tous les jours (Koskela and Väänänen-Vainio-Mattila, 2004; Mäyrä et al., 2006; Demiris et al., 2004; Kang et al., 2006; Callejas and López-Cózar, 2009). Les systèmes proposés envisagent d'apporter une assistance selon trois axes principaux :

– la *santé* lorsqu'il s'agit de suivre l'état de la personne et l'évolution de sa perte d'autonomie en employant des capteurs physiologiques, des détecteurs de mouvement, des caméras, etc. qui permettent de suivre l'évolution du poids, du rythme cardiaque, de l'activité ;
– la *sécurité* lorsqu'il s'agit de prévenir et de détecter des situations de détresse ou de danger grâce à des capteurs de chute, des détecteurs de fumée, d'intrusion dans le domicile, etc. ;
– et le *confort* pour les systèmes domotiques qui visent à compenser des handicaps grâce à un accès plus aisé aux appareillages domestiques.

Il convient de noter qu'un quatrième axe doit être ajouté, il s'agit de la *communication avec l'entourage et avec l'extérieur* qui est essentielle pour la personne isolée à domicile.

Dans tous les cas, un bâtiment ou habitat intelligent pour la santé sera équipé de capteurs qui permettront de suivre les activités de l'occupant du logement. Les information de ces capteurs seront analysées pour détecter la situation courante et pour apporter le retour ou l'assistance requise. Elles pourront aussi être utilisées pour détecter les activités de la vie quotidienne et contrôler l'évolution de l'autonomie de la personne suivie.

Naturellement, le bâtiment intelligent pose des questions d'éthique car les données recueillies, parfois des images, peuvent transiter par différents réseaux. Les études qui ont abordé ce problème l'ont fait par l'intermédiaire de questionnaires auprès de la population cible, c'est à dire les personnes âgées, la famille, les aidants et les personnels de soins (Rialle, 2007a; Beaudin et al., 2006; Rumeau et al., 2006). Il en ressort tout d'abord le constat qu'il y a une très grande variabilité selon l'âge et les pathologies éventuelles, le type des données recueillies, et que la technologie est peu acceptée si elle ne s'adapte pas à la pathologie et aux besoins de l'utilisateur.

1.4 Quelques projets marquants ayant pour but l'assistance à domicile

Le nombre de projets liés au bâtiment intelligent est devenu tellement important qu'il devient difficile d'en faire un tour complet ; c'est pourquoi nous nous limiterons à la présentation de quelques projets qui nous semblent les plus pertinents. Seul parmi ces projets, le projet CompanionAble, présenté en section 1.4.4, prend en compte l'information audio, alors que ce canal apporte une information qu'il est parfois impossible d'obtenir par ailleurs lorsqu'il s'agit d'un cri, ou qui présente une forte valeur sémantique lorsqu'il s'agit d'un ordre

vocal ou d'un appel à l'aide. Cette information est prise en compte par Brdiczka et al. (2009) et pour la parole par le projet HERMES (Jianmin Jiang and Zhang, 2009). Cet aspect fera l'objet du chapitre 2. Par ailleurs, les projets RESIDE-HIS(2000-2002) (Istrate et al., 2006) et DESDHIS(2002-2004) (Vacher et al., 2006) avaient pour but la conception, la mise au point et l'expérimentation d'un dispositif de télémédecine s'appuyant sur l'utilisation de capteurs sonores, de capteurs de déambulation et d'activité. Les buts recherchés étaient la reconnaissance des situations de détresse et le suivi d'activité.

1.4.1 House_n - Massachusetts Institute of Technology (MIT)

Le projet House_n [15] (Intille, 2002) couvre les domaines de l'informatique, de l'architecture, de la mécanique et de la médecine préventive. Il est dirigé par des chercheurs du département d'architecture qui s'occupent de la conception de la maison, des technologies et des services à intégrer afin que ceux-ci puissent évoluer pour faciliter les principales activités de la vie quotidienne, telles que la communication avec l'extérieur, l'assistance médicale, et, puissent améliorer l'autonomie des personnes âgés (Kent et al., 2003). Les objectifs principaux sont :

- l'application de techniques d'interaction homme-machine pour encourager des comportements sains (régime alimentaire, exercices...) ;
- la reconnaissance des activités — surtout pour les personnes âgées — ;
- et la surveillance biométrique à travers des dispositifs mesurant le rythme cardiaque, la tension artérielle, la respiration et la glycémie.

Une habitation, le *Place Lab*, a été complètement équipée afin d'étudier le comportement des personnes et leurs interactions avec les objets et l'environnement de l'habitat. Des volontaires ont accepté de vivre à l'intérieur pendant une période de quelques jours à une semaine. Une particularité de ce projet est le nombre impressionnant de capteurs utilisés pour enregistrer les activités développées (plusieurs centaines). Parmi les capteurs utilisés, on compte :

- des capteurs d'état sans fil sur les objets utilisés ou manipulés par les personnes, comme les contacts de portes, les fenêtres, les portes de commode, ainsi que dans les récipients de la cuisine ;
- des dispositifs à fréquence radio pour localiser les habitants ; des microphones pour capturer l'information audio ;
- un système de capture vidéo qui inclut des caméras infrarouges ;
- des dispositifs PocketPC pour recevoir les retours de l'utilisateur ;
- et des capteurs embarqués attachés sur le corps du volontaire pour la surveillance biométrique.

Il est important de noter que les volontaires ne sont pas spécifiquement des personnes âgées ou handicapées, mais des personnes de la population active.

15. http://architecture.mit.edu/house_n/

Une des préoccupations majeures du projet est la reconnaissance d'activités. Par exemple, ses membres préconisent une approche basée sur un réseau bayésien (Tapia et al., 2004) et affirment que l'utilisation d'un modèle dynamique tel qu'un HMM ne serait pas pertinente dans un environnement multicapteur. Les résultats expérimentaux montrent une précision entre 25% et 89% pour la reconnaissance (hors-ligne) de 35 activités possibles. En fait, le projet consiste surtout à tester de nouveaux capteurs et de nouvelles interfaces et à recueillir des données pour analyser le comportement de la personne. Le projet ne fournit pas, à notre connaissance, d'architecture logicielle globale pour permettre à l'habitation de capturer l'environnement, pour prendre des décisions, et pour agir sur l'environnement.

1.4.2 GER'HOME - CSTB

GER'HOME [16] (Zouba et al., 2009) est un projet du Centre Scientifique et Technique du bâtiment (CSTB) dont le but est d'appliquer des technologies de la domotique à l'aide au maintien à domicile des personnes âgées. Ce projet, maintenant terminé, s'est surtout concentré sur la reconnaissance des activités de la vie quotidienne (AVQ) à partir de l'analyse de vidéos et de traces de capteurs binaires dans un appartement dédié (contacts de porte, débitmètre...). La reconnaissance des activités se fait par un système expert qui utilise un langage de description d'événements (Van-Thinh et al., 2003) ; ce langage permet d'exprimer des contraintes spatio-temporeles entre les événements atomiques et certaines variables. Contrairement à la plupart des projets qui utilisent des outils de fouille de données pour acquérir la connaissance, ils utilisent une base de connaissances définie a priori par un expert qui inclut des modèles 3D pour la reconnaissance de formes (utilisés pour reconnaitre, par exemple la posture de l'habitant) et des descriptions sémantiques de l'environnement et des événements (par exemple, le fait qu'une personne s'assoit normalement sur une chaise et pas sur une table). Leurs résultats pour la reconnaissance des états et événements d'une personne âgée (64 ans) montrent une précision de 71-94% et une sensibilité de 62-87% (Zouba et al., 2009). Encore une fois, comme la plupart des projets, ce projet se concentre sur la reconnaissance de situations et la faisabilité technologique et n'inclut pas d'interaction avec l'utilisateur. L'architecture adoptée reste donc de type pipeline.

1.4.3 Aging In Place - University of Missouri

Aging in Place [17] (Skubic et al., 2009) est un projet qui a pour but de donner aux personnes âgées un espace indépendant qui comprend des capteurs dédiés afin de garantir une surveillance médicale adéquate et qui soit confortable de manière à ce que les habitants se sentent chez eux. Un bâtiment nommé *Tiger Place* (Rantz et al., 2008) a été construit respectant les normes sanitaires pour faire des expérimentations avec des volontaires qui ont occupé cet espace pendant des périodes allant de 3 mois à 3 ans (15 mois en moyenne). Parmi les volontaires, on trouve des personnes avec des maladies chroniques telles que l'arthrite,

16. http://gerhome.cstb.fr/
17. http://aginginplace.missouri.edu/

le diabète, des troubles cardiaques ainsi que des stades initiaux de la maladie d'Alzheimer. Divers professionnels (infirmiers, thérapeutes, physiologistes, ergonomes) ont participé à la création de ce bâtiment qui contient 31 appartements indépendants, mais aussi des espaces communs pour l'exercice physique et les loisirs. Chaque appartement est composé d'une cuisine, de toilettes, d'une chambre, d'une salle à manger et d'un séjour. L'originalité de ce projet est qu'il ne s'agit pas seulement d'installations ponctuelles mais de véritables logements qui ont été occupés par de véritables occupants pendant des périodes très longues.

Le principal objectif de *Tiger Place* est la détection de situations de détresse à travers l'analyse des données obtenues par les capteurs. Le système mis en œuvre *In-Home Monitoring System (IMS)*, comprend des dispositifs tels que des capteurs de mouvement et des contacts de porte (placard et réfrigérateur). Pour permettre un plus grand confort et éviter une trop grande intrusion, les dispositifs embarqués ont été évités. La figure 1.1 montre l'architecture du système. Comme dans d'autre approches (p.ex. : GER'HOME) les sources d'information sont analysées séparément puis fusionnées pour une meilleure reconnaissance de l'activité de la personne. Le module de raisonnement, basé sur de règles floues temporelles, utilise les données pour apprendre des motifs de comportement et inférer des changements importants. Les résultats sont stockés dans une base de données et peuvent être accédés via le web. Un module d'alerte a été mis en place de façon à ce que des événements d'intérêt puissent être notifiés lorsqu'ils se produisent. En ce qui concerne la reconnaissance vidéo, celle ci a été utilisée dans le but de compléter les informations des autres capteurs notamment pour la détection des événements anormaux, tels qu'une chute, et éviter les fausses alarmes. Le problème concernant le respect de la vie privée a été contourné en faisant une séparation entre le contour de l'habitant (qui apparaît comme un silhouette blanche) et l'arrière-plan. Les responsables du projet affirment que, bien que les personnes âgées ne soient pas à l'aise avec cette technologie vidéo dans leurs appartements, elles acceptent plus facilement l'utilisation de ce type d'images où seules des silhouettes grossières sont enregistrées. Pour l'extraction de ces silhouettes, des méthodes basées sur la logique floue ont été appliquées (Xi et al., 2006).

1.4.4 CompanionAble - UE

CompanionAble [18] est un projet financé par la communauté européenne. Son but est l'aide à la stimulation cognitive et à la gestion de la thérapie des personnes suivies à domicile. Ce support est donné à travers un « compagnon » robotique (mobile) fonctionnant de manière collaborative avec l'environnement informatique ambiant de l'habitat (statique). Le but et de construire un système d'assistance basée sur l'observation multimodale et l'interaction homme-machine. Un des aspects importants de ce projet est l'utilisation de la reconnaissance du son pour la détection de situations de détresse (Rougui et al., 2009), et l'utilisation de la reconnaissance de la parole pour le dialogue avec un robot (Caon et al., 2010). Le projet a démarré en 2008, et livre, pour l'instant peu d'informations sur son architecture

18. http://www.companionable.net/

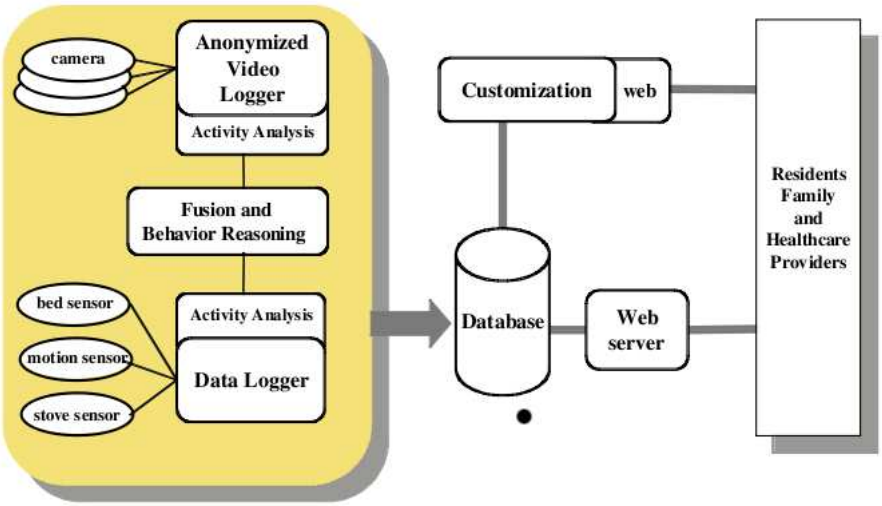

FIGURE 1.1 – Architecture du système In-Home Monitoring System

logicielle. Cependant, une originalité de ce projet consiste à inclure une interface audio pour la communication entre le robot et l'habitat. Par exemple, le système EMUTEM (Environnement Multimodal pour la Télévigilance Médicale) (Medjahed et al., 2009) a été développé dans le cadre de ce projet. Il est composé de trois sous-systèmes :

1. Anason, qui utilise un ensemble de microphones pour contrôler et surveiller l'environnement acoustique de l'habitant ;
2. RFpat, un système embarqué qui mesure le rythme cardiaque, la posture et éventuellement la chute de la personne ; et
3. Gardien, un ensemble de capteurs à infrarouge qui localise non seulement l'habitant mais identifie aussi sa posture.

La méthode de fusion de données obtenues à partir des trois sources est basée sur la logique floue, pour tirer partie des possibilités offertes par cette logique pour traiter l'imprécision. Pour chaque sous-système, un ensemble d'entrées a été défini, et, pour chacune des entrées, une liste d'ensembles flous. Par exemple, pour Anason, les entrées sont la classification de l'environnement selon le son et le niveau de Rapport Signal sur Bruit (RSB). Les listes d'ensembles flous correspondants sont : *no signal*, *normal*, *possible alarme*, *alarme* et *bas*, *moyen*, *haut*.

Les sorties du système sont des alarmes (avec les niveaux $normal$ et $situation\ de\ de-tresse$) et la localisation de l'habitant où les pièces sont affectées de niveaux flous. Le moteur d'inférence utilise deux groupes de règles. Le premier contrôle la sortie pour la localisation et utilise les données des capteurs infrarouge et le RSB des microphones, tandis que le deuxième utilise toutes les données pour contrôler les sorties d'alarme.

Le système a été testé avec des scénarios basés sur des situations réelles qui cherchent à refléter les situations des personnes âgées. Ils ont obtenu 97% de bonnes détections pour la sortie d'alarme, et 95% pour la localisation.

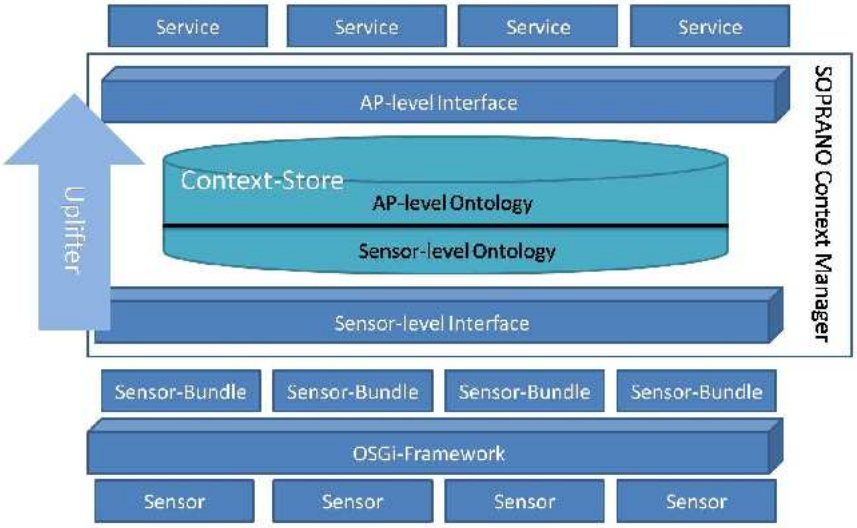

FIGURE 1.2 – Architecture de OpenAAL du projet SOPRANO

1.4.5 SOPRANO - UE

SOPRANO (*Service-Oriented Programmable Smart Environments for Older Europeans*) [19] (Klein et al., 2007) est un projet financé par la Commission Européenne qui vise à aider les personnes âgées à avoir une vie plus indépendante. L'apport principal du projet a été la création de l'intergiciel (*middleware*) OpenAAL pour la mise en œuvre logicielle des maisons intelligentes. Cette infrastructure logicielle ouverte prend comme entrée les informations provenant des capteurs installés dans la maison, et active grâce à ses sorties des services pour effectuer des actions sur l'environnement. La figure 1.2 montre l'architecture de OpenAAL. Elle est orientée vers les services (*Service Oriented Architecture*, SOA) et basée sur OSGi. L'originalité du système est de fournir deux couches pour les représentations sémantiques de l'environnement : la première est une description des capteurs et de leurs états et, la deuxième décrit la situation de l'habitant, ses activités, sa localisation, les urgences.

L'interface au niveau des capteurs permet d'inclure facilement des capteurs grâce à la définition d'un protocole de communication entre un capteur et l'interface. L'information fournie par les capteurs est décrite par une ontologie « métier » spécifique aux capteurs qui fournit un vocabulaire commun pour l'échange des données. Ainsi, les fournisseurs de capteurs peuvent intégrer leurs composants indépendamment des autres composants du système en respectant cette description. L'interface de service est mise en œuvre à un plus haut niveau. Elle permet aux services développés sur SOPRANO de faire des requêtes sur l'information contextuelle, telle que la localisation de l'habitant.

Un exemple d'un tel service est un logiciel détectant une situation de détresse spécifique et envoyant un appel d'urgence. La transformation de l'information pour passer du niveau capteur à celui de service est opérée par le composant *uplifter* qui implémente un algorithme de raisonnement ; cet algorithme est basé sur des HMM et des réseaux de neurones qui infèrent le contexte présent à partir de données incertaines et imprécises. Cependant,

19. http://www.soprano-ip.org/

cet algorithme n'est pas une partie intégrante du système, d'autres *uplifters* se basant sur des méthodes différentes peuvent être intégrés indépendamment des capteurs et services.

1.5 Positionnement par rapport à ces projets

Les projets Ger'Home et House_n ne prévoient pas d'interaction à l'initiative de l'utilisateur, ils visent essentiellement à suivre les activités de la personne dans son environnement au moyen de capteurs et à faire un suivi des Activités de la Vie Quotidienne. Le projet House_n prévoit aussi la mise en place de systèmes aidant la personne à avoir un comportement sain (activité physique, nourriture) et permettant d'estimer son état de santé par des capteurs physiologiques.

Le projet Soprano a défini une ontologie et une architecture logicielle de service basée sur OSGi. Les projets Aging in Place, Companionable et Soprano accordent une grande importance à la détection des situations de détresse. De plus le projet Companionable a pour objectif la stimulation cognitive et la gestion de la thérapie des personnes. Il envisage pour ce faire de développer un système de dialogue avec le robot incluant une commande vocale, mais il n'y a pas eu à notre connaissance de publication de résultats à ce sujet. En ce qui concerne l'analyse sonore, ce projet prévoit aussi la détection de parole et de sons de la vie courante dans l'appartement comme éléments d'information du système de génération d'alarme.

Ces projets visent donc essentiellement la détection de situation de détresse, le suivi de l'état de santé et de la prise de médicaments, et la stimulation cognitive. Ils s'intéressent donc aux deux aspects *Santé* et *Sécurité*, mais sans aborder les deux autres aspects essentiels qui sont le *Confort* et la *Communication avec l'entourage et avec l'extérieur*. En effet, les Maisons Intelligentes tendent à être équipées de dispositifs dotés d'interfaces de plus en plus complexes et en conséquence d'autant plus difficiles à maîtriser par l'usager. Les personnes qui bénéficieraient le plus de ces nouvelles technologies sont les personnes en perte d'autonomie telles que les personnes atteintes de handicaps moteurs ou fragilisées par des pathologies dues à l'âge (Alzheimer). Elles sont les moins aptes à utiliser des interfaces complexes, étant donné leur handicap ou leur manque de familiarité avec les nouvelles technologies. Il devient donc indispensable de prévoir une assistance facilitant la vie quotidienne et l'accès à l'ensemble des systèmes dits « domotiques ». Les interfaces tactiles usuelles devront être complétées par des interfaces plus accessibles, ne sollicitant ni la vue, ni le mouvement, grâce notamment à un système réactif à la parole. Ces nouvelles interfaces trouveront également leur utilité lorsque la personne ne peut plus se déplacer ou se déplace difficilement.

J'aborderai donc plus précisément dans ce mémoire l'apport d'informations d'origine sonore en complément d'autres sources pour la commande vocale, la localisation et la reconnaissance d'activités, les 2 derniers points constituant une information sur le contexte nécessaire avant la prise de décision par le système automatique. L'analyse sonore et la re-

connaissance de la parole dans un environnement réel présentent de nombreuses difficultés ; le chapitre 2 s'appuiera sur une expérience pendant laquelle des personnes ont joué un scénario dans un appartement pour mettre en évidence les défis scientifiques à relever dans ce domaine. Le chapitre 3 fera ensuite un état des lieux de l'utilisation de la commande vocale dans la Maison Intelligente et présentera une étude d'acceptabilité conduite auprès de personnes âgées, de leurs proches et de personnels médicaux.

CHAPITRE 2

Analyse d'informations sonores dans un bâtiment intelligent

L'analyse automatique des sons et de la parole concerne de nombreux domaines du fait de l'intérêt croissant porté aux systèmes de surveillance automatique. Les sons de la vie courante sont générés par les gestes quotidiens (vaisselle, clefs, porte...), ils sont désignés habituellement par le mot « bruit ». Après avoir décrit très brièvement l'état de l'art dans ce domaine, nous exposerons les méthodes que nous avons conçues et mises en œuvre pour l'analyse sonore en temps réel. Ensuite, nous décrirons une expérience et nous en tirerons les conclusions. Le lecteur pourra trouver plus de détails en se reportant à (Vacher et al., 2011) et aux autres articles de ma bibliographie personnelle listés en annexe (section 5 du curriculum vitae).

2.1 Analyse audio

L'analyse audio, et tout particulièrement la reconnaissance de la parole, est un domaine de recherche très ancien. De nombreuses méthodes, ainsi que des paramètres acoustiques très variés, ont été explorés et l'état de l'art des techniques est fortement relié à l'apprentissage automatique de modèles probabilistes (réseaux de Markov cachés, réseaux neuronaux...). Les récents développements ont donné des résultats significatifs et ont permis à la reconnaissance automatique de la parole d'être une composante de beaucoup de produits industriels, mais il subsiste de nombreux défis à relever pour rendre cette fonctionnalité disponible pour le bâtiment intelligent. L'analyse sonore peut être scindée en deux branches :

- la reconnaissance de la parole pour la commande vocale et le dialogue, et,
- la reconnaissance des sons pour identifier l'interaction de la personne avec son environnement (par exemple la fermeture d'une porte) ou avec le fonctionnement d'un appareil (par exemple d'une machine à laver).

En ce qui concerne l'identification des sons, certains projets visent à déterminer l'état de santé de la personne vivant dans un bâtiment intelligent au travers de la reconnaissance d'activités ou de la détection de situations de détresse. L'analyse sonore a été utilisée pour quantifier l'usage de l'eau (hygiène et boisson) (Ibartz et al., 2008) mais elle reste peu utilisée pour la détection de situations de détresse (Istrate et al., 2006). Malgré tout, Litvak et al. (2008) ont utilisé des microphones et un accéléromètre pour détecter une chute dans un appartement, alors que Popescu et al. (2008) ont utilisé 2 microphones dans le même but. Hors du contexte de détection de situation de détresses, Chen et al. (2005) ont déterminé

différentes activités dans une salle de bain en utilisant des HMM sur des coefficients cepstraux à fréquence mel (*Mel-Frequency Cepstral Coefficients*, MFCC). Cowling (2004) a utilisé la reconnaissance de sons non vocaux en combinaison avec leur direction d'origine pour un robot de surveillance autonome.

Les systèmes de Reconnaissance Automatique de la Parole (RAP), ou *Automatic Speech Recognizer* (ASR), ont atteint de bonnes performances lorsque le microphone est placé à proximité du locuteur. C'est par exemple le cas lorsqu'on utilise un casque audio, mais les performances se dégradent très vite lorsque le microphone est placé à plusieurs mètres, par exemple dans le plafond. Cela est dû à différents phénomènes comme la présence de bruit de fond et la réverbération. Ces problèmes doivent être pris en compte dans le contexte de l'assistance à domicile, certaines solutions seront présentées au chapitre 4.

2.2 Analyse audio en temps réel dans un habitat intelligent

Cette section présente le système d'analyse sonore temps réel AUDITHIS qui intègre les résultats antérieurs des études sur la détection et la classification des sons faites dans le cadre des projets RESIDE-HIS (IMAG) (Istrate et al., 2006) et DESDHIS (ACI Santé) (Vacher et al., 2006) qui avaient pour objectif la détection d'une situation de détresse de la personne. Ce système, qui sera présenté en section 2.2.3, a été utilisé dans un habitat intelligent pour la santé (HIS) au cours de deux expérimentations, pour l'évaluer en ce qui concerne la reconnaissance des appels de détresse et l'utilisation de la dimension sonore pour la classification des Activités de la Vie Quotidienne. Ces 2 expérimentations ont permis de mettre en évidence les défis restant à relever pour que l'analyse audio apporte des informations utiles dans le domaine de l'habitat intelligent.

2.2.1 Détection des sons

La détection constitue la première phase d'un système d'extraction d'informations par analyse sonore. Il s'agit de déterminer à chaque instant s'il y a présence ou absence de signal dans le bruit de fond, ce qui permet d'isoler un événement sonore, son de la vie courante ou parole, en vue d'un traitement ultérieur. Les signaux de parole et les signaux de son de la vie courante présentent des caractéristiques très différentes, ce qui fait que les différentes techniques classiques dites *Voice Activity Detection* (VAD) ne sont pas applicables (Beritelli et al., 2002; Nemer et al., 2001).

Une méthode de détection des événements sonores utilisant la transformée en ondelettes de Daubechies a été proposée (Vacher et al., 2004). La méthode s'appuie sur la détermination de l'énergie des plus fort coefficients d'ondelette en utilisant un seuil adaptatif. Cette méthode a été validée sur le bruit blanc et sur des bruits enregistrés dans un appartement expérimental, et donne de bons résultats avec un rapport signal sur bruit (RSB) égal à 0 dB (Istrate et al., 2006) pour lequel les méthodes de reconnaissance et de classification ne peuvent conduire à des résultats satisfaisants.

TABLE 2.1 – Corpus de texte utilisé pour générer un modèle de langage

Catégorie	Exemple	Corpus d'origine	Nombre	Total
Détresse	"Au secours", Un médecin vite"	Anodin/Détresse	60	
		Complément	33	
				93
Conversation courante	"Bonjour", "Où est le sel", "J'ai bu ma tisane"	Anodin/Détresse	66	
		Complément	172	
				238
Ordre domotique	"Monte la température"	Complément		39
Toutes				415

2.2.2 Traitement des évènements sonores

Séparation son/parole

Lorsqu'un événement sonore a été détecté et extrait, il est tout d'abord nécessaire de déterminer s'il s'agit d'un son de la vie courante ou de parole, puisque les traitements diffèreront suivant le cas. Le problème consiste à classifier le signal comme parole ou comme son de la vie courante, et on a prouvé que les méthodes classiques GMM ou HMM sont applicables pour cela (Vacher et al., 2007).

Classification des sons

Plusieurs classes de sons correspondant au problème posé ont été identifiées, soit parce qu'ils sont en rapport avec une activité de la personne (claquement de porte, serrure de porte, sonnerie de téléphone, sons de pas, vaisselle), soit parce qu'ils sont en rapport avec un état possible de détresse (bris de verre, chute d'objet, cris). Un corpus a été enregistré et a permis d'apprendre des modèles GMM (Istrate et al., 2006) ou HMM (Vacher et al., 2007). Les résultats ont été très satisfaisants avec au taux d'erreur de classification égal à 12,2% en utilisant des coefficients cepstraux à fréquence linéaire (*Linear Frequency Cepstral Coefficients*, LFCC).

Reconnaissance des mots de détresse

La base du système de reconnaissance est le système RAPHAEL (Vaufreydaz et al., 2000), construit autour du moteur Janus (Kratt et al., 2004) développé par le Karlsruhe Institute of Technology et le CMU (Carnegie Mellon University), qui analyse le signal de parole et fournit une meilleure hypothèse de reconnaissance à partir de laquelle il sera possible ou non d'identifier des mots clefs de détresse. Ce système de reconnaissance automatique de la parole (RAP) comprend un étage de reconnaissance phonétique utilisant des modèles acoustiques, un étage de reconnaissance de mots construisant le graphe de phonèmes et utilisant un dictionnaire phonétique, et un étage de reconnaissance de phrases utilisant un modèle de langage.

L'apprentissage des modèles acoustiques a été réalisé sur de grands corpus pour assurer une bonne indépendance par rapport au locuteur. Ces corpus ont été recueillis par environ 300 locuteurs de langue française aux laboratoires CLIPS (BRAF100) et LIMSI (BREF80 et

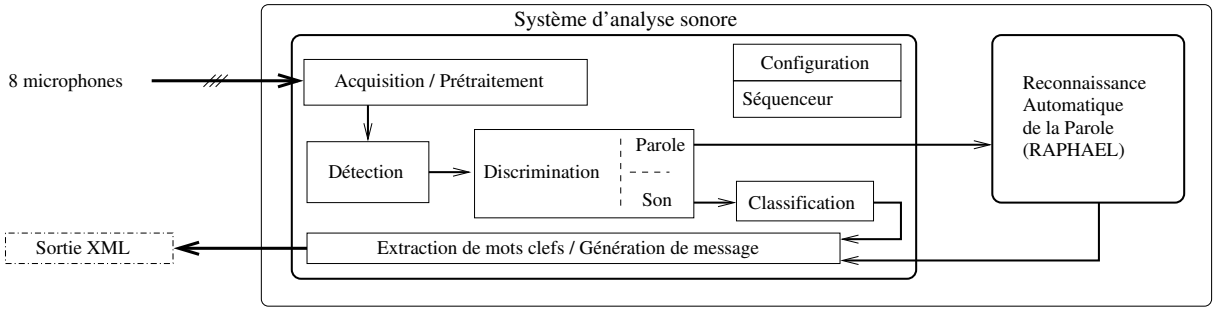

FIGURE 2.1 – Architecture du système AUDITHIS

BREF120). Chaque phonème est modélisé par un modèle de Markov à 3 états (HMM). Le signal est découpé en trames de 16 ms avec un recoupement de 50%.

L'exigence est de parvenir le mieux possible à la détection de situations de détresse grâce à la détection de mots-clefs, sans nécessairement reconnaître l'ensemble de la conversation. Pour ce faire, nous avons utilisé les phrases du corpus de parole Anodin/Détresse (Vacher et al., 2006) auxquelles ont été ajoutés des ordres domotiques (”monte la température”...) et d'autres phrases. Le corpus Anodin/Détresse (38 minutes, 2646 phrases) a été enregistré au laboratoire par 21 locuteurs (11 hommes et 10 femmes) entre 20 et 65 ans, il comprend 66 phrases caractéristiques d'une situation normale de la personne, et 60 phrases de détresse. Comme le montre le tableau 2.1, l'ensemble comprend donc un total de 370 phrases (39 ordres domotiques, 93 phrases de détresse et 238 phrases anodines). Cet ensemble de phrases a permis de générer un modèle de langage à très petit vocabulaire (299 unigrammes, 729 bigrammes et 862 trigrammes) qui ne peut permettre la reconnaissance que d'un nombre très réduit de phrases, pour ne pas risquer de porter atteinte à l'intimité de la personne.

2.2.3 Le système AUDITHIS

L'organisation générale du système d'analyse sonore AuditHIS est présentée à la figure 2.1. Il recueille en temps réel le signal fourni par 8 microphones qui peuvent être des microphones omnidirectionnels sans fil (Sennheiser eW500), plus faciles à installer dans un appartement que des microphones filaires. Le système tourne dans un environnement GNU/Linux. Chaque événement sonore est recueilli et traité séparément par les différents modules :

- le module d'acquisition et de prétraitement qui procède à l'acquisition du signal sur chacune des voies de la carte multicanal et à l'estimation du niveau de bruit ;
- le module de détection qui détermine les débuts et fins des événements sonore, et peut évaluer le RSB associé à chaque événement ;
- le module de discrimination parole/son qui aiguille le signal vers le classifieur ou le système de reconnaissance de la parole ;
- le classifieur GMM ou HMM et le système Raphael qui déterminent la classe de son ou la parole la plus probable ;
- le module d'extraction de mots-clefs et de génération de messages qui enregistre l'en-

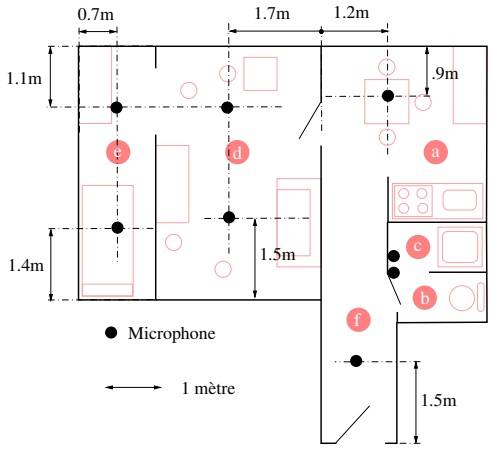

FIGURE 2.2 – Plan de l'appartement HIS du bâtiment Jean Roget (Faculté de Médecine de Grenoble) et disposition des microphones

semble des informations relatives à chaque événement sonore (date, durée, RSB, présence d'événements simultanés).

Les différents modules sont des processus légers qui s'exécutent séparément de manière autonome.

2.3 Expérimentation dans un appartement

2.3.1 Conditions expérimentales dans l'habitat intelligent

Les expérimentations ont été réalisées en conditions réelles dans le HIS (Habitat Intelligent pour la Santé) du laboratoire TIMC-IMAG (Le Bellego et al., 2006). Cet appartement de 45 m^2 est équipé de différents types de capteurs dont 7 microphones que nous utiliserons pour l'analyse sonore. Il comprend : (a) une cuisine, (b) des toilettes, (c) une salle de bains, (d) un séjour, (e) une chambre, et, (f) un couloir. La figure 2.2 montre la disposition des microphones qui sont installés dans chaque pièce au plafond et dirigés vers le sol. Cet appartement constitue un environnement difficile du point de vue sonore, d'une part du fait de la présence possible et imprévisible de bruit extérieur (hélicoptère arrivant à l'hôpital et monte-charge à proximité immédiate), et, d'autre part à cause de la présence de zones vitrées face à face dans la chambre et le séjour qui sont sources de réverbération. Les participants à l'expérimentation ont pu à chaque fois se placer comme ils le souhaitaient dans l'appartement sans tenir compte de leur position par rapport à celle des microphones.

2.3.2 Détection d'un appel de détresse

Pour évaluer le système, chaque participant a suivi un scénario comportant la prononciation de 45 phrases du corpus Anodin/Détresse (20 phrases de détresse, 10 phrases anodines et 3 conversations téléphoniques de 5 phrases chacune). Dix personnes (3 femmes et 7 hommes), âgés de 37,2±14 ans, ont participé à cette expérience.

Les phrases ont été prononcées par la personne assise ou debout dans la pièce ; la distance de la personne aux microphones placés dans la même pièce était donc de 1 à 10 mètre. Le téléphone était placé sur une table dans le séjour.

La personne devait entrer dans l'appartement, effectuer quelques activités dans la cuisine et le couloir, puis rentrer dans le séjour en fermant la porte derrière elle. Elle devait alors aller dans la chambre et prononcer la moitié des phrases anodines et des phrases de détresse, puis revenir dans le séjour et prononcer le reste des phrases. Le téléphone sonnait et une des conversations téléphoniques prévues s'engageait. Ce dernier processus était répété encore 2 fois.

Les données ayant un RSB supérieur à 5 dB ont été traitées au fil de l'eau, mais les fichiers contenant les enregistrements et leurs données XML ont été conservés pour constituer un corpus. Ce seuil a été fixé de manière empirique à partir des résultats de la classification sur des données bruitées (Vacher et al., 2006). Au total, en éliminant les occurrences simultanées de moindre RSB et les signaux saturés, et, après annotation manuelle, le corpus d'étude est composé de $nDS = 197$ phrases de détresse et de $nNS = 232$ phrases anodines.

Ces 429 phrases ont ensuite été traitées par RAPHAEL en utilisant les modèles acoustiques et les modèles de langage présentés précédemment. Les performances ont été évaluées à partir du Taux d'Alarmes Manquées (TAM), du Taux de Fausses Alarmes (TFA) et du Taux Global d'Erreur (TGE) définis de la manière suivante :

$$\begin{cases} TAM = \frac{nDM}{nDS} \\ TFA = \frac{nFA}{nNS} \\ TGE = \frac{nDM+nFA}{nDS+nNS} \end{cases} \tag{2.1}$$

en considérant le nombre de Fausses Alarmes nFA, et, le nombre de Détections Manquées nDM. Il y a Fausse Alarme lorsque il y a un mot-clef de détresse dans l'hypothèse de reconnaissance d'une phrase anodine, et, Détection Manquée lorsqu'il n'y a pas de mot-clef de détresse dans l'hypothèse de reconnaissance d'une phrase de détresse.

Les résultats sont présentés pour chaque locuteur dans le tableau 2.2. Globalement, le taux de Fausses Alarmes est assez bas $TFA = 4\%$ quel que soit le locuteur. Par contre le Taux d'Alarmes Manquées est important $TAM = 29,5\%$, et varie fortement selon le locuteur, entre 5% et 55%. Les plus mauvaises performances ont été observées avec un locuteur qui prononçait les paroles de détresse comme un acteur de tragédie ($TAM = 55\%$ et $TFA = 4,5\%$). Les variations d'intensité on fait que le pronom « je » n'était pas audible en début de phrase. Un autre locuteur s'exprimait en marchant avec des talons ; le bruit des talons se superposant à la parole a entraîné une mauvaise reconnaissance en masquant certains phonèmes de la phrase. Une des phrases de détresse est constituée du mot « help », qui a été mal reconnu lorsque les locuteurs l'ont prononcé en aspirant fortement, car le phonème [h] n'existe pas en français et ne faisait donc pas partie des modèles acoustiques. Par ailleurs, lorsque une phrase débute par un claquement de langue ou un bruit environnemental, le premier phonème reconnu est préférentiellement une fricative ou une occlusive, ce qui altère la reconnaissance.

TABLE 2.2 – Taux d'erreur de reconnaissance de mots-clef caractéristiques d'une alarme

Locuteur	TAM (%)	TFA (%)	TGE (%)
1	19.0	0.0	8.9
2	5.3	0.0	2.3
3	35.0	8.7	21.0
4	16.7	4.2	9.5
5	55.0	4.5	28.6
6	36.8	4.2	18.6
7	23.5	4.3	12.5
8	42.9	5.0	24.4
9	33.3	0.0	15.5
10	24.0	8.7	16.0
Performance globale	29.5	4.0	15.6

Cette expérimentation montre malgré tout une forte dépendance du système de reconnaissance vis à vis du locuteur, ce qui devrait être corrigé par une adaptation du système au locuteur, ainsi que l'importance des perturbations apportées par le bruit et les conditions expérimentales. Les performances peuvent cependant être améliorées en utilisant le fait qu'on enregistre simultanément sur plusieurs canaux (Vacher et al., 2008).

2.3.3 Analyse audio dans les activités de la vie quotidienne

Une expérimentation a été montée pour tester le système AUDITHIS en conditions semi-dirigées. Pour que les données soient acquises dans les conditions les plus réalistes possibles, on a demandé aux participants de réaliser au moins une fois 7 activités (AVQ) dans l'appartement. Ces activités inclaient :

1. le sommeil,
2. le loisir : regarder la télé, écouter la radio, lire un magazine...
3. l'habillement,
4. la prise de nourriture : réaliser et prendre un repas,
5. l'élimination en allant aux toilettes,
6. l'hygiène : se laver les dents et les mains... et,
7. la communication avec l'extérieur : utiliser le téléphone.

De cette manière, l'expérimentation nous a permis d'obtenir des événements sonores représentatifs et directement en relation avec les activités journalières usuelles. Il serait fort intéressant de permettre par analyse sonore de contribuer à la détermination du degré d'autonomie de la personne en fonction de l'échelle de Katz (Katz and Akpom, 1976), ce qui n'a pas été possible car les performances des systèmes de classification et de reconnaissance n'ont pas été suffisantes, notamment parce qu'il n'existait pas de modèle pour toutes les catégories de sons émises lors de l'expérimentation. Cet aspect sera étudié en section 2.4.

Classe	Son de la vie courante	Parole
Son de la vie courante	1745 (87%)	252 (23%)
Parole	141 (25%)	417 (75%)

TABLE 2.3 – Matrice de confusion Parole/Son de la vie courante

Quinze personnes ont participé successivement à l'expérimentation et ont réalisé les 7 activités sans qu'il y ait de condition sur le temps passé. La moyenne d'âge était 32±9 ans (l'âge variant entre 24 et 43 ans) et la durée d'une expérience est allée de 23 min 11 s à 1h 35 min 44 s. Une visite préalable de l'appartement a permis de montrer où l'on pouvait trouver tout ce qui était nécessaire à la réalisation des activités. Chaque personne était libre de réaliser les activités dans l'ordre de sa préférence. Plus de détails sur cette expérimentation sont décrits dans (Fleury et al., 2010a).

La discrimination entre son et parole est importante, d'une part car les traitements ultérieurs diffèrent, mais aussi d'autre part à cause de l'interprétation différente qui en sera faite par exemple pour la reconnaissance d'activité. Les résultats de la discrimination entre parole et sons de la vie courante pour les 2555 événements sonores enregistrés sont donnés dans le tableau 2.3 et montrent un taux d'erreur d'environ 25%. Par exemple, les sons de vaisselle ont été souvent confondus avec la parole du fait que le corpus d'apprentissage n'était pas assez riche, et aussi à cause de la présence d'une fréquence fondamentale proche de celle de la parole. Les cris diffèrent assez peu de la parole ; il faut distinguer le cas des cris articulés comme « Aïe » qui pourront être reconnus par le système de RAP et le cas des autres cris comme « Aaaaa » qui doivent plutôt être considérés comme non verbaux.

L'ensemble des sons recueillis constitue en soi un corpus riche en enseignements. Au total 1886 sons individuels et 669 signaux de parole ont été analysées et annotées manuellement. La durée totale est seulement de 8 minutes et 23 secondes, ce qui peut paraître peu mais il faut tenir compte du fait qu'un son dure en moyenne 0,19 seconde et que les personnes ont seulement prononcé des phrases courtes au téléphone. La vie quotidienne génère des sons d'une grande diversité, et les classes de sons sont beaucoup plus nombreuses que celles prévues initialement, ce qui explique en partie les mauvais résultats de classification.

Parmi les sons non prévus, on peut relever ceux produit par l'arrachement d'un adhésif Velcro, le tonnerre, la pluie sur les vitres... Le tableau 2.4 montre la répartition des sons en grandes catégories. Le détail des classes de sons du corpus est disponible dans (Vacher et al., 2011). La première catégorie est celle des sons humains (toux, raclement de gorge, sifflement, chant...) qui sont en relation avec la personne dans la pièce. La deuxième catégorie est celle de la manipulation d'objets et de fournitures ; elle est fortement reliée à l'activité de la personne. Une catégorie particulière est l'eau courante qui peut apporter des informations intéressantes sur les activités de préparation des repas, d'hygiène ou d'élimination. Une partie importante des sons n'a pu être identifiée, soit parce que leur source elle-même n'était pas identifiable, soit parce que trop de sons étaient enregistrés au même moment. Il est important de relever que certains sons ont été produits à l'extérieur du logement et que

les sons non identifiés représentent plus de 22% de la durée du corpus.

Classe	Taille de la classe	RSB moyen (dB)	Durée moyenne (ms)	Durée totale (s)
Sons humains	36	12,1	101	3,4
Manipulation d'objets	1302	11,9	59	76,3
Sons extérieurs	45	9,0	174	7,9
Appareillage domestique	72	8,0	209	15,1
Eau courante	36	10,1	1756	63,2
Divers	395	9,5	94	37,1
Ensemble des sons (hormis la parole)	1886	11,2	108	203,3

TABLE 2.4 – Regroupement des sons en grandes catégories

2.4 Analyse sonore pour l'assistance à domicile et défis à surmonter

L'analyse sonore (de sons et de parole) présente un potentiel important pour le suivi de la personne, la compensation de handicaps, et l'assistance, au travers de la commande vocale pour le confort et de l'appel à l'aide pour la sécurité. Elle peut aussi être utile pour améliorer et faciliter la communication de la personne vis à vis de l'extérieur, et peut aussi jouer un rôle dans le suivi d'activités en vue d'une évaluation de l'autonomie de la personne.

Cependant, ainsi que l'expérimentation précédente l'a confirmé, les technologies (reconnaissance de parole, dialogue, synthèse de parole, détection sonore) doivent prendre en compte le milieu difficile de l'habitat. Dans cette partie, nous allons aborder les applications possibles et les principaux défis à relever.

Extraction de l'information audio dans un environnement bruité

En conditions réelles, le signal audio est souvent perturbé : 1) par la présence d'un bruit de fond indésirable (TV, appareils divers, circulation routière) ; 2) l'acoustique de la pièce ; 3) la position du locuteur vis à vis du microphone. Le niveau de bruit du corpus recueilli varie entre 5 et 15 dB, ce qui est très éloigné des conditions de studio et explique la différence des résultats de reconnaissance. Les sons étaient d'une très grande diversité, alors que les personnes agissaient dans un cadre assez contraint malgré la liberté dans la façon de réaliser leurs activités. Le niveau de signal était variable suivant la position dans la pièce ; le fait d'être assis augmente aussi la distance vers les microphones qui sont fixés au plafond. La présence de très grandes surfaces vitrées en vis à vis a entraîné des altérations du signal, mais ce qui s'avère le plus gênant est la présence de signaux simultanés, due à des sources parasites comme le poste de TV. Certaines techniques de traitement du signal prenant en compte simultanément l'ensemble ou plusieurs des microphones devront être considérées (séparation de sources aveugles, analyse en composante principale, *beam-forming*...).

La présence d'un environnement bruité est fréquente et très perturbante pour les systèmes de reconnaissance, c'est pourquoi nous approfondirons cet aspect au chapitre 4.

Suivi d'activité et état de santé

Une application utile peut être la reconnaissance du fonctionnement des appareils domestiques, des écoulements d'eau pour évaluer la façon dont la personne utilise son environnement pour effectuer ses activités.

Dans une moindre mesure, la fréquence de détection de sons comme la toux, le raclement de gorge et les reniflements peut constituer un indicateur de problèmes de santé. Une indication de la solitude de la personne peut être déduite en utilisant la reconnaissance de la parole d'autres personnes ou la détection de l'usage du téléphone. Pour une application domotique, la détection du son peut s'avérer une source de localisation utile pour agir en fonction de la localisation de la personne en vue de son confort (éclairage qui suit le déplacement de la personne la nuit…). Une application plus ambitieuse pourrait être d'identifier la communication non verbale pour évaluer le bien-être ou la souffrance d'une personne démente.

Évaluation de la capacité à communiquer

La reconnaissance de la parole pourrait jouer un rôle important pour le suivi de personnes démentes. En effet, l'un des symptômes les plus tragiques de la maladie d'Alzheimer est la perte progressive de vocabulaire et de la capacité à communiquer. Un suivi constant de l'activité de la personne pourrait permettre de détecter des phases importantes dans l'évolution de la démence. L'analyse audio est la seule modalité qui offre la possibilité d'un suivi automatique de la perte de vocabulaire, de la décroissance des périodes de prise de parole, de l'isolement dans la conversation, etc. Les changements peuvent être très lents et difficiles à détecter par le personnel soignant.

Interface vocale pour la compensation et le confort

L'application la plus directe est la possibilité d'interagir verbalement avec l'environnement de la maison intelligente au travers d'une commande vocale ou d'un dialogue pour procurer un confort aux personnes handicapées ou fragiles. Cela peut aussi être fait indirectement avec coupure de la lumière lorsque la personne quitte une pièce. Un tel dispositif peut être important pour la prise en compte d'un handicap. De plus, la reconnaissance d'un jeu de commandes réduit est plus facile que la reconnaissance de l'ensemble des phrases que peut prononcer une personne.

Détection de situations de détresse

L'identification des sons de la vie courante peut être particulièrement intéressante pour l'évaluation d'une situation de détresse dans laquelle la personne peut se trouver. Par exemple, le bris de glace est couramment utilisé pour les alarmes. En outre, dans le cas où la personne

est restée consciente mais ne peut plus bouger, par exemple suite à une chute, le système offre la possibilité d'appeler à l'aide par la voix.

Reconnaissance des sons de la vie courante

En général, on considère deux catégories principales : les sons de la vie courante et la parole. Ces 2 catégories représentent une information sémantique complètement différente et les techniques de reconnaissance utilisées diffèrent. La reconnaissance des sons de la vie courante est un domaine assez récent, et, les paramètres acoustiques aussi bien que les classifieurs ne sont pas standardisés (Dufaux, 2001; Cowling, 2004; Istrate et al., 2006; Fleury et al., 2010a; Tran and Li, 2009). Les auteurs présentent dans la plupart des cas des méthodes probabilistes avec des modèles appris sur des corpus et visent à identifier les sons d'après leur source. Or, l'expérimentation menée dans le HIS a montré qu'il y avait une très grande diversité des sons obtenus et donc beaucoup de classes. En outre, dans la plupart des cas, la durée totale de chaque classe est très faible, soit parce que la classe comprend très peu d'éléments, soit parce que la durée de chacun d'eux est aussi très faible. Il est donc très difficile d'acquérir un corpus qui permettrait d'obtenir une classification correcte des sons dans une maison intelligente. Par contre, une méthode de classification hiérarchique s'appuyant sur les caractéristiques intrinsèques du signal (périodicité, fréquence fondamentale, forme temporelle impulsive ou large, croissante ou décroissante...) pourrait être une solution pour améliorer le système en nécessitant peu d'apprentissage. Une autre possibilité consisterait à lever des ambiguïtés en utilisant d'autres sources d'information disponibles dans le logement pour déterminer un contexte courant. Le système intelligent de supervision pourrait alors utiliser ces informations pour associer l'événement sonore à une source d'émission et prendre une décision adaptée à l'application.

Reconnaissance de la parole adaptée au locuteur

Le public concerné est très varié mais les personnes âgées en constituent une part appréciable. Plusieurs expérimentations effectuées en reconnaissance automatique de la parole ont montré des dégradations de performance pour les populations dites « atypiques » telles que les enfants ou les personnes âgées (Wilpon and Jacobsen, 1996; Vipperla et al., 2008; Gerosa et al., 2009) et l'intérêt d'une adaptation aux populations visées (Gerosa et al., 2009). D'autres études plus générales (Vipperla et al., 2010; Gorham-Rowan and Laures-Gore, 2006) mettent en évidence les effets du vieillissement sur la production de la parole et les conséquences que cela implique sur la reconnaissance de parole. Les locuteurs âgés se caractérisent par des tremblements de la voix, des hésitations, une production imprécise des consonnes, une cassure de la voix, et une articulation plus lente. La reconnaissance de la parole adaptée à la voix des personnes âgées est un domaine encore peu exploré, notamment en français. Nous nous y sommes attaqués depuis quelque temps (Dugheanu, 2011).

Vie privée et acceptabilité

Il est important de rappeler que la reconnaissance de la parole doit respecter la vie privée du locuteur. À cet égard, le modèle de langage doit être adapté à l'application et ne doit pas permettre de reconnaître des phrases dont le sens n'est pas indispensable à l'application ; un système de reconnaissance de mot-clefs respecte cette contrainte. Un système utilisé en contexte réel doit reconnaitre la parole au fil de l'eau pour la reconnaissance d'activité ou d'appel de détresse et ne doit pas permettre une reconnaissance ultérieure détaillée.

Un autre aspect concernant l'acceptabilité du système est le fait qu'un système sera d'autant plus accepté que la personne pourra effectivement se rendre compte que le système lui est utile la plupart du temps plutôt que pendant un instant éphémère (comme par exemple lors d'une chute). Un système global (surveillance, domotique et détection de détresse) sera plus facilement accepté.

Le respect de la vie privée et l'acceptabilité sont deux aspects déterminants. C'est pour cette raison qu'une étude spécifique sur l'usage et l'acceptabilité d'une interface vocale pour les personnes âgées a été conduite. Cette étude ainsi que ses résultats seront présentés dans le chapitre 3.

CHAPITRE 3

Acceptabilité de l'interface vocale dans l'habitat intelligent

3.1 État de l'art de l'utilisation de la reconnaissance de la parole dans la maison intelligente

Un certain nombre d'études récentes ont été conduites sur l'utilisation des technologies audio dans les habitats intelligents. Elles ont porté sur la reconnaissance de la parole (Vacher et al., 2010), la reconnaissance des sons (Rougui et al., 2009; Vacher et al., 2010), la synthèse de la parole (Lines and Hone, 2006) et le dialogue (Hamill et al., 2009; López-Cózar and Callejas, 2010). Des opinions concernant l'intérêt des personnes âgées en ce qui concerne l'interaction vocale ont été collectées par Callejas and López-Cózar (2009); Koskela and Väänänen-Vainio-Mattila (2004). Dans une étude récente, des personnes ont montré leur intérêt pour une commande vocale permettant de fermer les fenêtres et les rideaux ou de mettre en route la télévision ou la radio (Callejas and López-Cózar, 2009), et 95% de ces personnes ont déclaré qu'elles continueraient à utiliser le système même si il faisait parfois des erreurs en interprétant les ordres. Dans (Koskela and Väänänen-Vainio-Mattila, 2004), l'ordre vocal est utilisé pour interagir pendant la réalisation de petites tâches à la cuisine, par exemple pour répondre au téléphone tout en cuisinant. Les entretiens ont mis en évidence la crainte des personnes que le système ne reconnaisse pas ce qu'elles disent, ce qui, d'après les auteurs, pourrait être dû à leur expérience des interfaces de téléphones mobiles. Toutes ces études ont montré que la voix reste une modalité prometteuse pour améliorer la sécurité, le confort et l'assistance dans les maisons intelligentes, mais qu'elle reste encore peu explorée en comparaison avec les modalités classiques (interrupteur, commande à distance, PDA, téléphone mobile).

Un résultat commun de toutes ces études est que, quelle que soit la technologie audio considérée, aucune application relative à la maison intelligente ne peut rencontrer de succès si les utilisateurs potentiels ne sont pas inclus dans le processus de conception du système (Fugger et al., 2007; Mäyrä et al., 2006; Callejas and López-Cózar, 2009; Augusto, 2009). L'acceptabilité est un facteur clef pour intégrer de nouvelles technologies chez les particuliers, surtout s'il s'agit de personnes âgées ou ayant peu de formation dans les technologies informatiques. Les technologies d'assistance seraient donc développées en vain s'il n'y avait pas de réelle prise en compte des besoins, des craintes et des aspirations des personnes.

3.2 Démarche suivie

L'étude présentée a été réalisée dans le cadre du projet SWEET-HOME financé par l'Agence Nationale de la Recherche (ANR) 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 (Sweet-Home, 2010). Une présentation plus complète de cette étude est disponible dans (Portet et al., 2011).

L'*acceptabilité* est un critère fortement lié à l'utilisateur humain qui pourra ainsi être intégré tout au long de la conception grâce à une approche centrée sur la personne (Jaimes et al., 2006).

L'implication d'utilisateurs futurs est indispensable car les possibilités des systèmes ubiquitaires dans les maisons intelligentes sont encore mal définies (Fugger et al., 2007).

La première étape a consisté à définir les spécifications du système à partir de l'expertise des partenaires du projet en maison intelligente, en assistance à domicile et en équipement social pour les anciens et aussi à partir de données de la littérature (Rialle et al., 2008; Callejas and López-Cózar, 2009; Edwards and Grinter, 2001; Koskela and Väänänen-Vainio-Mattila, 2004; Demiris et al., 2004; Kang et al., 2006; Mäyrä et al., 2006; Fugger et al., 2007). Les fonctionnalités et les méthodes d'interaction ont pu être ainsi précisées avant de confronter des utilisateurs potentiels à un système qu'ils croient automatique mais qui est en fait actionné par un Magicien d'Oz (Jambon, 2009). Cela a permis de recevoir des retours et des suggestions d'utilisateurs obtenus dans des conditions réalistes, qui ont ensuite été intégrés dans la conception. Suite à cette étape de Magicien d'Oz, les différentes fonctionnalités du système SWEET-HOME seront développées dans une démarche centrée sur le système, avant d'être intégrées avec l'environnement domotique pour des tests impliquant des utilisateurs en vue d'adaptations et corrections finales.

3.3 Conception expérimentale

Les 18 participants ont été recrutés dans la région grenobloise ; ils se répartissaient en 3 groupes : 8 personnes âgées, 7 proches (famille ou amis) et 3 professionnels de santé. La moyenne d'âge du premier groupe était de 79 ans, 5 sur 8 étaient des femmes. Il s'agissait de personnes autonomes vivant seules, une seule disposant d'un ordinateur. La moyenne d'âge du deuxième groupe était de 41 ans.

L'expérimentation s'est déroulée dans l'appartement DOMUS du Centre des Technologies Logicielles décrit dans (Portet et al., 2011) ; cet appartement fonctionnel de 34 m^2 comprend une cuisine/salle à manger, une chambre à coucher, une salle de bains et un bureau. Il est équipé de capteurs et d'actionneurs pour agir sur l'environnement. Il a aussi été équipé de 7 microphones sans fil placés dans le faux plafond. La figure 3.1 décrit l'emplacement des différents capteurs. Des caméras accrochées au plafond de toutes les pièces, sauf la salle de bains, permettent de suivre les expériences à partir de la régie. Les caméras et les actionneurs ont permis la mise en place du système de Magicien d'Oz qui a été piloté par un opérateur

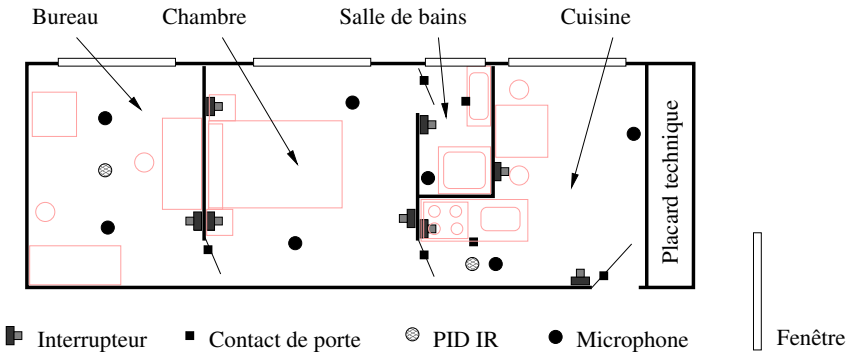

FIGURE 3.1 – Plan de l'appartement DOMUS du bâtiment DOMUS et disposition des capteurs

depuis la régie.

Chaque test a impliqué un ergonome qui a été en prise directe avec les personnes interrogées, un magicien d'Oz en régie pour piloter le système domotique à distance, et un couple composé d'une personne âgée accompagnée d'un de ses proches (sauf pour une personne qui est venue seule). L'ergonome a choisi une approche en codécouverte qui a permis à chaque personne accompagnée de son proche de prendre en compte le contexte expérimental tout en échangeant des points de vue avec son proche. L'acceptabilité du système a été recherchée en abordant les aspects suivants : l'utilité, l'ergonomie, la personnalisation, la manière d'interagir, l'intervention du système à la place de la personne, l'interruption d'une activité de la personne par le système, le maintien du lien social et la sécurité. Quatre scénarios ont permis à la personne d'être mise en situation de se servir de la commande vocale, de la communication avec l'extérieur, d'un agenda partagé, et aussi d'être interrompue par le système suite à une erreur.

- La personne âgée accompagnée de son proche était présente dans l'appartement, l'ergonome lui demandait de commander à la voix les rideaux, la lumière et la machine à café sans précision sur le mode opératoire. Le processus était ensuite réitéré en tenant en main une télécommande. À chaque fois, le magicien d'Oz simulait une mauvaise compréhension et ne déclenchait la commande qu'à la deuxième tentative. Après cela, un entretien permettait à la personne et à son proche de répondre séparément au sujet de leurs préférences.
- La personne âgée regardait seule la télévision dans le bureau lorsque soudain l'émission était interrompue, et son proche qui était dans la régie entrait en téléconférence avec elle. Ensuite son proche regagnait le bureau afin de répondre aux questions sur les préférences des deux personnes ;
- La personne et son proche discutaient dans l'appartement quand une voix préenregistrée était émise sur un haut-parleur leur rappelant de fermer la porte d'entrée ou d'arrêter la plaque électrique. Cela permettait de recueillir leurs réactions sur l'acceptabilité d'une interruption par le système, puis sur la sécurité.
- Après explications sur le fonctionnement d'un agenda électronique partagé, la personne et son proche étaient interrogés pour savoir s'ils accepteraient un tel dispositif

et quelles seraient les personnes autorisées à y accéder.

3.4 Résultats

Cette étude présente des biais, car elle a été conduite avec peu de personnes, qui sont toutes originaires de la même région et n'ont pas utilisé le système en conditions réelles. Cependant, il s'agissait surtout de mettre en évidence les freins et objections des utilisateurs du système testé, et l'approche par magicien d'Oz qui a permis de simuler le contexte de l'application est difficile à utiliser sur une grande population. L'expérimentation en codécouverte impliquant à la fois la personne âgée et son proche a permis à ces deux personnes de se projeter dans la situation et de confronter les deux points de vue qui diffèrent pour ces deux catégories.

Tous les participants on montré de l'intérêt pour une interaction vocale, ce qui confirme les études précédentes (Callejas and López-Cózar, 2009). Cependant, 2 personnes sur 8 ont émis des craintes en ce qui concerne l'enregistrement par microphone. Il faudra donc que les phrases recueillies soient effacées après traitement, ainsi que les transcriptions qui ne sont pas des ordres domotiques. L'intérêt manifesté pour la commande vocale est conforté par d'autres études qui ont montré que les systèmes d'urgence utilisant un bouton poussoir étaient peu adaptés (Hamill et al., 2009).

La commande par mot-clé est par ailleurs bien acceptée, ce qui facilite la réalisation du système de reconnaissance. Par ailleurs, les personnes se disent prêtes à utiliser le système même s'il peut faire des erreurs en interprétant les ordres.

En ce qui concerne les informations sonores émises par le système, la moitié des personnes préfère une voix féminine, une personne une voix masculine, et pour les autres cela est indifférent. Il faut noter que tous refusent une voix synthétique, comme l'avaient relevé Lines and Hone (2006). Les professionnels de santé ont trouvé intéressant de pouvoir recevoir des informations n'importe où dans la maison quand on a les mains et les yeux occupés.

Les personnes âgées et leurs proches ont exprimé leur crainte d'une trop forte dépendance de la personne vis à vis d'un système, ce qui fait que, dans le cas où le système tomberait en panne, la personne seule ne serait plus en mesure de s'adapter à cette situation.

Sur un autre aspect, les personnes souhaitent impérativement conserver leurs activités quotidiennes et maintenir le plus d'autonomie possible, un système leur faisant gagner du temps n'est pas nécessairement adapté à leurs souhaits. Par ailleurs, les professionnels de santé craignent que le système ne conduise lentement à une limitation des visites des proches qui se contenteraient de la visioconférence : dans les faits, cela renforcerait le sentiment d'isolement de la personne.

La sécurité est la plus forte attente. Le système proposé peut éviter à la personne de se mettre en danger en se déplaçant rapidement pour trouver son téléphone, ou en ne trouvant pas l'interrupteur la nuit ; il peut aussi alerter lorsque la porte est restée ouverte. Cet aspect est confirmé par un certain nombre d'études sur la sécurité des personnes âgées (Rantz et al., 2008; Callejas and López-Cózar, 2009; Rialle et al., 2008). Les personnes ont aussi exprimé

leur crainte qu'un intrus ne puisse commander le système par la voix à leur place.

La vie privée doit être respectée, et, alors que peu de personnes âgées ont déclaré être effrayées par un enregistrement sonore, presque toutes refusent l'utilisation d'une caméra (7/8), ce qui est une confirmation de l'étude de Demiris et al. (2004) qui ont interrogé 15 personnes de plus de 65 ans sur les habitats intelligents. Rialle (2007a) a mis en évidence que l'acceptabilité d'une technologie dépend fortement de la détresse et de la dépendance de la personne, mais les avis que nous avons recueillis concernent des personnes autonomes.

Dans le projet SWEET-HOME, une caméra est utilisée pour la téléconférence avec les proches, elle doit être placée là où elle ne peut poser de problème. Peu d'études concernent les microphones ; les proches et les professionnels interrogés ont précisé qu'une interface vocale ne leur paraissait pas une atteinte à l'intimité. Par contre, il peut y avoir sentiment d'intrusion lorsque le système intervient pour demander une confirmation : un indicatif bref (son ou musique) doit permettre de rassurer la personne qui pourrait penser que quelqu'un s'est introduit dans l'appartement.

Le calendrier partagé est unanimement refusé : en effet, les informations ne sont pas de même nature suivant les catégories de personnes et présentent très souvent un aspect confidentiel qui s'oppose à leur divulgation.

L'objection principale exprimée par les personnes pendant l'expérimentation est que le système, quoique très intéressant, n'est pas adapté à leur état présent, étant donné qu'elles ne sont pas dépendantes. Elles ont du mal à se projeter dans le futur, mais reconnaissent que le système conviendrait pour des personnes dépendantes. Elles insistent aussi sur le fait que le système doit les assister mais ne doit pas diminuer leur autonomie. Elles vivent seules et elles pensent que continuer à effectuer des activités banales et sans danger (faire le café, tirer les rideaux, etc.) représente un aspect important de leur vie quotidienne et leur permet de rester en forme et autonomes. Ce dernier point est confirmé par d'autres études à l'étranger (Callejas and López-Cózar, 2009; Augusto, 2009). Le maintien du lien social par le biais de la visioconférence sur la télévision proposé dans le système SWEET-HOME est un aspect important qui a soulevé l'intérêt des personnes interrogées. Cependant, les professionnels de santé ont souligné que cela ne doit pas remplacer de vraies visites.

CHAPITRE 4

Reconnaissance de la parole en conditions distantes

L'un des plus grands problèmes qui freinent le développement des techniques basées sur la parole dans les logements réels est la faible performance des systèmes de reconnaissance de la parole dans cet environnement, qui est en général bruité (circulation automobile à l'extérieur, appareils domestiques, etc.), avec de la réverbération, et où les microphones ne sont pas nécessairement placés à proximité immédiate du locuteur. Tous ces problèmes qui sont relatifs à la *parole distante* doivent être pris en compte dans le cadre de l'habitat intelligent.

Les applications usuelles des systèmes de RAP se placent dans le cas où le microphone est placé à quelques centimètres de la bouche du locuteur. Lorsque les microphones sont placés au plafond et dirigés vers le sol, il n'y a plus un trajet unique de la source vers le récepteur mais une combinaison de trajets qui font intervenir l'acoustique de la salle. L'adaptation des systèmes de reconnaissance à ce type de signaux peut se faire soit au niveau des modèles acoustiques, soit par l'utilisation de paramètres acoustiques adaptés (Wölfel and McDonough, 2009). Il a par ailleurs été montré par (Deng et al., 2000) que l'utilisation de paramètres adaptés procure de meilleures performances que celles obtenues par des systèmes entraînés avec des données ayant une distorsion comparables à celle des données à reconnaître. Par ailleurs, lorsque le temps de réverbération reste inférieur à 500 millisecondes, ce qui est le cas d'un logement, les performances ne sont pas améliorées de manière significative lorsque l'on utilise des données d'apprentissage enregistrées dans les mêmes conditions de réverbération (Baba et al., 2002).

La section 4.1 qui suit est consacrée à la fusion de données fournies par un système de RAP disposant d'enregistrements recueillis par plusieurs microphones placés dans un appartement, en s'intéressant plus particulièrement aux ordres domotiques et aux appels de détresse prononcés par une personne. Par ailleurs, la présence de bruit additionnel constituant une autre difficulté rencontrée en reconnaissance de la *parole distante*, la section 4.2 traitera de l'annulation du bruit lorsque la source de bruit est identifiée.

4.1 Reconnaissance de la parole à partir de plusieurs canaux

Lorsque plusieurs microphones sont utilisés simultanément, l'information que l'on cherche à extraire est présente sur plusieurs voies et devrait donc permettre d'obtenir de meilleures performances que celle obtenue à partir d'une voie unique. Il est possible de reconnaître séparément la parole prononcée sur chacune des voix avant de faire une fusion de données

Phase 1	40 énoncés	862 phrases	Durée par canal : 38 minutes 46s
Phase 2	44 énoncés	917 phrases	Durée par canal : 40 minutes 27s

TABLE 4.1 – Caractéristiques du corpus de parole

avec un vote majoritaire (Vacher et al., 2008) ou un algorithme ROVER ; il est possible aussi d'utiliser des méthodes plus en amont comme le décodage guidé (Lecouteux et al., 2008). Cette section correspond à un travail mené en collaboration avec Benjamin Lecouteux (Lecouteux et al., 2011a,b).

4.1.1 Corpus de test et système de reconnaissance

Un corpus de test a été acquis avec les 7 microphones disposés dans l'appartement DOMUS décrit précédemment au chapitre 3 (voir figure 3.1). 21 personnes de 22 à 63 ans dont 7 femmes ont participé à l'enregistrement d'un corpus multimodal dans un contexte proche de la vie quotidienne à partir d'un scénario en 2 phases. Le sous-corpus de parole utilisé pour les tests a été extrait de ce corpus avant d'être annoté. Le rapport signal sur bruit a été estimé pour chaque phrase sur chacune des 7 voies. Les caractéristiques principales du corpus sont résumées dans la table 4.1. La phase 1 est composée de 40 énoncés prononcés au cours de conversations téléphoniques, la phase 2 comprend 44 énoncés dont 10 indiquant un appel à l'aide. Le corpus est composé de 862 phrases pour la phase 1 (38 minutes 46s par canal) et pour la phase 2 de 917 phrases (40 minutes 27s par canal). Le rapport signal sur bruit moyen est de 20,3 dB sur le meilleur canal.

Le système de RAP Speeral (Linarès et al., 2007) a été choisi après des tests impliquant plusieurs autres systèmes, et aussi parce qu'il implémente le décodage guidé. Pour rester proche d'une application temps réel, c'est la configuration 1xRT qui a été utilisée. Speeral s'appuie sur un décodeur A* avec des modèles acoustiques dépendant du contexte basés sur des HMM et des modèles de langage à base de trigrammes. Les HMM utilisent un modèle classique à 3 états gauche-droite, et, l'obtention des états utilise des arbres de décision. Les paramètres acoustiques sont composés de 12 coefficients PLP (*Perceptual Linear Predictive*), l'énergie ainsi que les dérivées premières et secondes de ces 13 paramètres.

Les modèles acoustiques de départ ont été appris sur le corpus ESTER contenant environ 80 heures de parole annotée. Ils ont ensuite été adaptés à chacun des 21 locuteurs en utilisant la méthode MLLR (*Maximum Likelihood Linear Regression*) et les données correspondantes annotées de la phase 1. Le modèle de langage comporte un lexique de 10 000 mots. Il a été obtenu par l'interpolation d'un modèle générique (pondération 10%) et d'un modèle spécialisé (pondération 90%). Le modèle générique était estimé à partir de corpus en français (Le Monde et Gigaword) et de transcriptions de nouvelles fournies pendant la campagne ESTER. Le modèle spécialisé était estimé à partir de l'ensemble de toutes les phrases prononcées par les locuteurs pendant les phases 1 et 2 (ordres domotiques, phrases anodines ou de détresse).

4.1.2 Utilisation de la méthode ROVER

La méthode ROVER (*Recognizer Output Voting Error Reduction*) (Fiscus, 1997) est destinée à améliorer les résultats de reconnaissance en harmonisant au mieux les sources les plus fiables. Pour cela, elle combine les sorties des systèmes en un réseau unique de transitions entre mots. Ensuite, un vote est effectué au niveau de chaque point de branchement et le mot choisi est celui auquel est affecté le meilleur score (nombre de votes pondérés par leur niveau de confiance). Cette méthode est assez gourmande en calcul, car elle nécessite pour nous d'effectuer la reconnaissance sur les 7 canaux, ou au moins sur les deux canaux ayant le meilleur rapport signal sur bruit.

Lorsqu'un signal $s(t)$ est mélangé à un bruit $b(t)$, le signal observé $e(t)$ est $e(t) = s(t) + b(t)$, et le rapport signal sur bruit sur la durée ΔT est donné par la relation :

$$RSB(e) = 10 * \log\left(\frac{\Sigma_{t\in[0;\Delta T]} s^2(t)}{\Sigma_{t\in[0;\Delta T]} b^2(t)}\right) \tag{4.1}$$

Souvent, seul le signal $e(t)$ est accessible ; le signal $b(t)$ n'est donc mesurable qu'en l'absence du signal $s(t)$ sur un autre intervalle de temps. L'équation 4.1.2 est remplacée par la formule approchée suivante qui suppose que le bruit est stationnaire et reste négligeable devant le niveau du signal :

$$RSB(e) = 10 * \log\left(\frac{\Sigma_{t\in[t_s;t_s+\Delta T]} e^2(t)}{\Sigma_{t\in[t_b;t_b+\Delta T]} b^2(t)}\right) \tag{4.2}$$

Chaque parole $P_{k,j}$ du canal j annotée est composée d'une partie signal de parole de durée I_{parole} entourée par deux zones de silence de durée $I_{silence-d}$ au début et $I_{silence-f}$ à la fin de durée totale $I_{silence} = I_{silence-d} + I_{silence-f}$. Le signal à chaque instant est $s(n)$ et le rapport signal sur bruit de la phrase considérée est défini par la relation :

$$RSB(P_{k,j}) = 10 * \log\left(\frac{\frac{\Sigma_{n\in I_{parole}} s(n)^2}{|I_{parole}|}}{\frac{\Sigma_{n\in I_{silence}} s(n)^2}{|I_{silence}|}}\right) \tag{4.3}$$

Le décodage sera en principe d'autant plus sûr que le RSB est élevé. Une mesure de confiance $\phi_{k,j}$ associée à une phrase $P_{k,j}$ pourra être choisie de la manière suivante :

$$\phi_{k,j} = \frac{2^{RSB(P_{k,j})}}{\Sigma_{i=1}^{7} 2^{RSB(P_{k,i})}} \tag{4.4}$$

Cette mesure de confiance permet de pondérer le vote lors du processus ROVER. L'erreur obtenue avec ces méthodes est exprimée en Taux d'Erreur de Mots (TEM) ; elle est donnée dans le tableau 4.2 accompagnée de son écart-type. La méthode ROVER pondérée améliore significativement les performances en utilisant seulement les 2 meilleurs canaux, puisque l'erreur qui était de 18,3% est ramenée à 13%. La fiabilité de cette méthode est confirmée par la diminution importante de l'écart-type.

L'utilisation de 7 canaux augmente beaucoup le temps de calcul, mais n'apporte pas

	Reconnaissance sur le canal de meilleur RSB	ROVER sur 7 canaux	ROVER sur 2 canaux
$TEM_{\pm\sigma}(\%)$	$18,3_{\pm 12,1}$	$12,2_{\pm 6,1}$	$13,0_{\pm 6,6}$

TABLE 4.2 – TEM pour la reconnaissance avec l'algorithme ROVER

d'amélioration significative ; elle présente donc peu d'intérêt.

4.1.3 Utilisation du décodage guidé

Le décodage guidé, ou DDA (*Driven Decoding Algorithm*), peut être utilisé au niveau de l'étage de décodage du système de RAP pour améliorer les performances en utilisant des transcriptions auxilliaires (Lecouteux et al., 2006, 2008). Le système de RAP génère des hypothèses au fur et à mesure qu'il progresse dans le treillis de phonèmes et le DDA agit à chacune de ces étapes. L'hypothèse courante est alignée avec la transcription auxilliaire et un score de correspondance α est calculé puis intégré au modèle de langage :

$$\tilde{P}(w_i|w_{i-1},w_{i-2}) = P^{1-\alpha}(w_i|w_{i-1},w_{i-2}) \tag{4.5}$$

en notant $\tilde{P}(w_i|w_{i-1},w_{i-2})$ la probabilité modifiée du trigramme connaissant le mot w_i et les précédents w_{i-1}, w_{i-2}, et $P(w_i|w_{i-1},w_{i-2})$ la probabilité initiale du trigramme.

Il est possible d'utiliser la sortie de décodage d'un premier canal de microphone comme entrée auxiliaire pour guider la reconnaissance d'un deuxième canal. Cela offre comme avantage d'une part d'augmenter la vitesse de décodage du deuxième système de RAP (seulement 0,1xRT au lieu de 1xRT), et, d'autre part de réunir véritablement et facilement les flux, alors que les stratégies de vote comme ROVER ne réunissent pas les sorties des systèmes de RAP. Le premier canal utilisé est pour chaque phrase le canal présentant le meilleur RSB et le deuxième canal celui présentant le second meilleur RSB. Cette première méthode sera désignée par DDA_{+RSB}.

Une deuxième méthode permet d'intégrer une connaissance *a priori* sur les phrases comportant des ordres domotiques, ce qui permettra d'améliorer les performances seulement en ce qui concerne ces phrases. Le système de RAP est guidé par les ordres domotiques reconnus lors de la première passe. Cette méthode sera désignée par $DDA_{2-level+RSB}$. Les segments de parole résultant de la première passe sont projetés sur les *3-meilleurs* ordres vocaux en utilisant une distance d'édition, et ensuite injectés dans le système de RAP lors de la deuxième passe. Pour localiser des ordres domotiques dans des transcriptions (treillis de phonèmes) T de taille m, chaque phrase de taille n d'ordre domotique H est alignée sur T en utilisant un algorithme de DTW (*Dynamic Time Warping*) au niveau phonétique (Berndt and Clifford, 1994). Les coûts de suppression, insertion et substitution sont calculés de manière empirique. La distance cumulée $\gamma(i,j)$) entre H_j et T_i est obtenue selon la relation :

$$\gamma(i,j) = d(T_i,H_j) + min\{\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\} \tag{4.6}$$

	Reconnaissance sur le canal de meilleur RSB	DDA_{+RSB}	$DDA_{2-level+RSB}$
$TEM_{\pm\sigma}(\%)$	$18,3_{\pm 12,1}$	$11,4_{\pm 5,6}$	$8,8_{\pm 3,7}$

TABLE 4.3 – TEM pour la reconnaissance utilisant un décodage guidé

L'erreur obtenue suivant les différentes méthodes, exprimée en taux d'erreur de mots (TEM), est donnée dans le tableau 4.3 en comparaison avec les performances de décodage obtenues sur le canal présentant le meilleur rapport sur bruit. Elle est accompagnée de son écart-type. Le décodage guidé à deux niveaux $DDA_{2-level+RSB}$ améliore significativement les performances, avec $TEM = 8,8\%$, par rapport à la reconnaissance sur le meilleur canal ($TEM = 18,3\%$), il convient aussi de noter la diminution notable de l'écart-type qui passe de $12,1$ à $3,7$. La méthode de décodage guidé DDA_{+RSB} utilise moins d'informations *a priori* et conduit naturellement à de moins bonnes performances ($TEM = 11,4\%$).

Avec le guidage par les ordres domotiques, les taux d'erreur sont moins importants que pour les méthodes utilisant un vote : $8,8\%$ contre $12,2\%$. Bien entendu, lorsqu'il n'y a pas d'ordre domotique, il y a moins de différence avec les méthodes utilisant un vote et la valeur obtenue est $11,4\%$. Le décodage guidé présente aussi l'avantage de nécessiter bien moins de temps de calcul. Malgré tout, les résultats obtenus avec chacune de ces deux méthodes montrent que la combinaison des informations obtenues par plusieurs microphones conduit à de meilleurs résultats que l'utilisation du microphone procurant le meilleur rapport signal sur bruit.

4.2 Reconnaissance en présence de sources de bruit

Le son émis par une source fixe comme la télévision ou la radio se superpose à la parole et rend la reconnaissance plus difficile. Il est possible de diminuer l'influence de cette source en utilisant un algorithme d'annulation d'écho, comme cela a été mis en œuvre dans (Vacher et al., 2009).

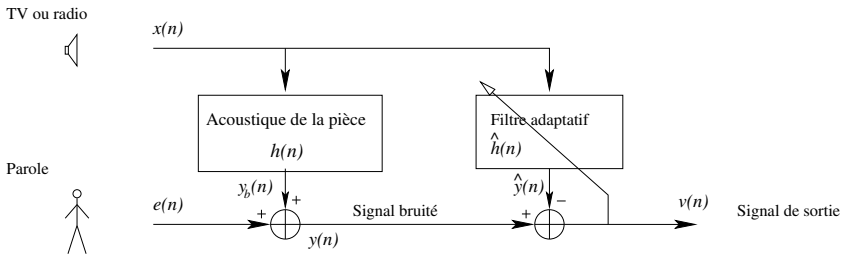

FIGURE 4.1 – Schéma de principe d'un système d'annulation d'écho

Le bruit $y(n)$ recueilli au niveau du microphone enregistrant la parole est altéré par l'acoustique de la pièce. On a $y_b(n) = h(n) * x(n)$, en notant $h(n)$ la fonction de transfert de la pièce, $x(n)$ le signal de bruit, et $*$ l'opérateur de convolution. Ce signal se superpose au signal de parole $e(n)$ au niveau du microphone, qui recueille alors le signal $y = e(n) + h(n) * x(n)$. Le

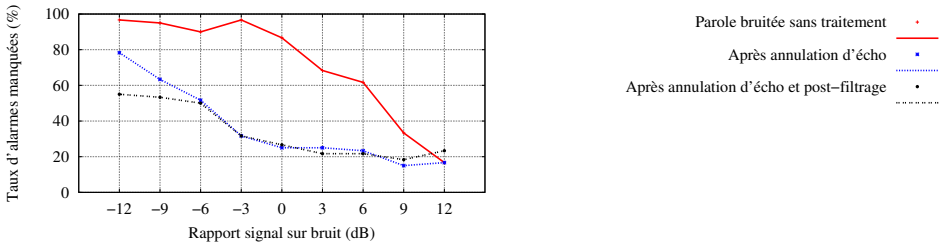

FIGURE 4.2 – Taux d'alarmes manquées en cas de bruitage par de la parole

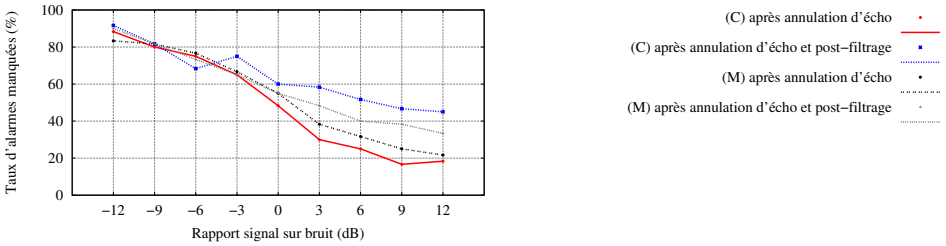

FIGURE 4.3 – Taux d'alarmes manquées en cas de bruitage par de la musique classique (C) et moderne (M)

principe de l'annulation d'écho consiste à estimer au moyen d'un filtre adaptatif le signal à retrancher pour annuler le bruit, comme indiqué à la figure 4.1. Le signal résultant est $v(n) = e(n) + h(n) * x(n) - \hat{h}(n) * x(n) = e(n) + \varepsilon(n)$.

La librairie SPEEX proposée par Valin (2007), basée sur une variante de l'algorithme *Normalized Least Mean Square* (NLMS), permet d'estimer la fonction de transfert $\hat{h}(n)$ en temps réel, et propose une méthode pour éviter que l'algorithme ne diverge en présence du signal $e(n)$ (cas appelé *double-talk* en téléphonie). Cette librairie a été intégrée à AUDITHIS pour des évaluations ; elle comprend aussi un étage de post-filtrage destiné à réduire le bruit introduit par l'étage d'annulation d'écho pour qu'il ne soit pas désagréable à l'oreille.

Une expérimentation a été réalisée dans une pièce du laboratoire pour enregistrer de manière synchrone le signal d'une radio $x(n)$ et le signal $y_b(n)$ après transformation par l'acoustique de la pièce au niveau d'un microphone placé dans la pièce à environ 4 mètres des haut-parleurs. Il a été ensuite possible d'ajouter au signal $y_b(n)$ les phrases du corpus Anodin-détresse pour obtenir le signal $y(n)$. Le signal $\hat{h}(n)$ peut ensuite être obtenu et soustrait à $y(n)$. L'intérêt de cette méthode est qu'on reste complètement maître du rapport signal sur bruit $RSB = \frac{e(n)}{y_b(n)}$ défini selon l'équation 4.1.2.

Il est alors possible d'étudier le comportement du système en fonction du rapport signal sur bruit pour plusieurs niveaux de mélange, par exemple entre $-12dB$ et $+12dB$. Dans cette expérience, l'acoustique de la pièce intervient au niveau du signal parasite $y_b(n)$ et non du signal $e(n)$ qui a été enregistré dans les conditions habituelles avec le micro proche de la bouche.

Les résultats ont été évalués par le taux d'alarmes manquées (TMA) en sortie du décodeur RAPHAEL présenté en section 2.2.2 à la page 33. Le modèle de langage choisi était un modèle statistique de taille moyenne (9958 mots) obtenu à partir du journal Le Monde

interpolé avec le texte du corpus Anodin/détresse. La taille du tampon de l'algorithme d'annulation d'écho était de 256 échantillons pour optimiser le temps de calcul ; la taille du filtre était de 8192 échantillons, c'est à dire suffisamment grande pour tenir compte de la taille de la pièce et des retards après réverbération sur les murs et la fenêtres.

Le système a été testé avec une émission d'actualités comportant essentiellement de la parole (France-Info) et deux sortes de musique : « classique » (3^{e} symphonie, opus 55 de L. van Beethoven) et d'avant-garde (Artificial Animals Riding on Neverland de AaRON). Les figures 4.2 et 4.3 montrent les résultats respectifs. Les résultats observés sont meilleurs dans le cas du bruitage avec de la parole : le taux d'erreur varie peu entre 0 et 12dB, même s'il y a une légère dégradation entraînée par l'annulation d'écho à 12dB par rapport à la reconnaissance sur le signal non débruité.

En ce qui concerne la musique, le post-filtrage utile pour l'auditeur humain entraîne de fortes détériorations. La musique moderne est plus difficile à annuler du fait de la présence de sources ayant une grande bande spectrale comme les percussions. Pour la musique, les performances sont nettement moins bonnes : on obtient les mêmes performances à 6dB (classique) ou 9dB (moderne) que celles obtenues à 0dB dans le cas du bruitage par de la parole.

En présence de bruits de provenances diverses, le problème est encore plus compliqué, car il n'est pas possible d'avoir accès au bruit à sa source. Les techniques proposées se limitent alors à la séparation de sources aveugles. L'analyse en composantes indépendantes peut s'adapter au cas simple du mélange linéaire entre le signal et le bruit, mais cela n'est pas suffisant pour modéliser les phénomènes rencontrés dans la réalité, où peuvent être présentes des composantes non linéaires ou convolutives. D'autres solutions comme l'adaptation du modèle acoustique au bruit sont envisageables, mais ne peuvent être mises en œuvre que lorsque le niveau de bruit reste modéré.

CHAPITRE 5

Localisation d'habitant par propagation d'activations multisources

5.1 Introduction

L'objectif du travail présenté dans ce chapitre est de définir une méthode de fusion de données pour la localisation de l'habitant à l'intérieur de son domicile en utilisant des sources d'informations non visuelles (c.-à-d. sans caméra), événementielles et indirectes (c.-à-d. sans capteur porté par la personne). Il se déroule dans le cadre du projet SWEET-HOME qui vise à concevoir un système de contrôle intelligent de la domotique à travers une interface de commande utilisant la voix, pour un plus grand confort et une plus grande sécurité de la personne. La localisation de la personne est primordiale pour interpréter les commandes vocales. Par exemple, si la personne prononce « allume la lumière », il est nécessaire de localiser la personne pour déduire la lampe à activer.

Les approches présentées dans la littérature reposent souvent sur l'utilisation de la vidéo (Friedland et al., 2010). Cependant, pour limiter les coûts et pour respecter la vie privée, seuls les capteurs domotiques sont considérés dans notre projet. Ces sources (ici capteurs de présence, capteurs de contact, et microphones) sont donc ambiguës (par exemple, observation d'un même évènement par plusieurs capteurs éloignés), ont une dimension temporelle (par exemple, dans le couloir à `15:23:45`) et dépendent de l'espace physique (p.ex., nombre de pièces, réverbération).

Parmi les capteurs considérés, le microphone est un capteur d'événements et de localisation prometteur qui, par sa nature omnidirectionnelle ou unidirectionnelle, peut être très sensible ou très spécifique. Cependant, très peu d'approches du domaine utilisent uniquement l'information sonore. On peut citer Bian et al. (2005) qui ont développé une grille de microphones dans une pièce pour déterminer l'origine d'un son dans un espace 3D avec une erreur moyenne de moins de 27 cm. Comme souligné par les auteurs, l'information audio nécessite beaucoup moins de bande passante que l'information vidéo et permet de détecter facilement certaines activités (conversations, sonneries de téléphone). Néanmoins, cette solution est trop complexe et trop coûteuse à mettre en place dans un appartement classique. De plus, si la vidéo est sensible aux changements de luminosité, le canal audio est sensible au bruit environnemental (Vacher et al., 2011). Le canal audio, bien qu'étant une modalité pertinente et peu coûteuse, est donc une source bruitée et ambiguë. Il devient donc essentiel de

mettre en place un système de localisation multisource qui tirerait partie des redondances et des complémentarités des sources choisies. Plusieurs travaux ont été menés dans ce sens, en utilisant principalement des méthodes de classification probabilistes telles que les réseaux bayésiens (Dalal et al., 2005) ou les modèles de Markov (Kröse et al., 2008).

Cependant, l'information temporelle dans ces modèles est assez pauvre et rarement prise en considération. Duong et al. (2009) prennent en compte l'aspect temporel à travers un modèle semi-markovien caché hiérarchique (c'est à dire prenant aussi en compte la durée dans l'état courant pour le changement d'état) pour la reconnaissance d'activités à partir de séquences d'événements en modélisant aussi le temps passé dans chaque état (représentant une sous-activité).

Cependant, dans notre cas, il est difficile de prédire la localisation présente en évaluant une séquence de localisations passées, car le nombre de chemins de déplacement possibles augmenterait l'incertitude. De plus, ces approches nécessitent un grand nombre de données étiquetées pour l'apprentissage.

Certaines applications mettant en jeu des traitements sur la connaissance, telles que la recherche d'information (Crestani, 1997; Aswath et al., 2005), la reconnaissance de formes (Niessen et al., 2008) ou la navigation d'agents animés (Kortenkamp and Chown, 1993), emploient une approche de propagation d'activation à travers un réseau sémantique. Ces réseaux sont constitués de nœuds représentant des concepts et de liens représentant l'intensité de la relation entre concepts. L'activation d'un nœud se dilue de proche en proche en activant les nœuds voisins jusqu'à ce qu'elle s'épuise. Des travaux récents (Niessen et al., 2008) ont montré l'intérêt de la propagation d'activation comme mémoire à court terme pour désambiguïser des événements sonores.

L'approche que nous avons adoptée pour localiser une personne se base sur un réseau dynamique où les événements des capteurs activent des hypothèses de localisation qui se périment au cours du temps. Ces hypothèses prennent en compte les activations précédentes, l'incertitude des événements et la durée de validité d'un événement. Dans notre approche, le réseau est dynamique dans le sens où il évolue en fonction des données, et où il prend en compte des connaissances a priori sur l'environnement dans lequel évolue l'habitant. Cette approche est détaillée dans la section 5.2. La méthode proposée est ensuite évaluée dans la section 5.3 sur des données réelles. Le chapitre se termine par une brève discussion.

5.2 Localisation d'habitant par propagation d'activations multisource

La méthode mise en œuvre pour la localisation d'une personne à partir de plusieurs sources d'information non visuelles repose sur la modélisation des liens entre les événements et les hypothèses de localisation par un réseau dynamique à deux niveaux. Chaque nouvelle information est propagée dans le réseau pour mettre à jour les hypothèses de loca-

lisation. Dans la plupart des approches de propagation d'activation, le réseau est bien établi, mais, dans notre cas, les informations ont une validité temporaire et les liens entre les nœuds évoluent en fonction du temps. La section 5.2.1 présente comment le réseau évolue au fil du temps en intégrant les données de plusieurs sources temporelles et comment cette information se « périme » grâce à une fonction d'oubli. La section 5.2.2 détaille comment des connaissances a priori sur l'environnement peuvent être prise en compte avec l'information immédiate des capteurs pour calculer les relations entre les différents nœuds du réseau.

5.2.1 Réseaux dynamiques et propagation d'activation

La méthode que nous utilisons est basée sur celle de Niessen et al. (2008) qui ont présenté une approche basée sur les réseaux dynamiques permettant de désambiguïser la reconnaissance d'événements sonores. Il s'agit d'un réseau à deux niveaux. Le niveau zéro est constitué des événements sonores, le niveau 1 représente les hypothèses liées à un évènement (par exemple, rebond de ballon ou claquement de mains), et le niveau 2 représente le contexte de l'évènement (par exemple, match de basket, concert). Chaque événement active des hypothèses selon l'évènement et les contextes auxquels les hypothèses sont liées. L'activation des hypothèses se propage ensuite aux contextes.

Nous avons adapté cette approche pour localiser une personne à partir de différentes sources d'événements. Le réseau dynamique que nous avons conçu est organisé en deux niveaux : le premier niveau correspond aux *hypothèses de localisation* générées à partir d'un *événement*, et le deuxième niveau représente les *contextes d'occupation* de chaque pièce ; le poids d'activation des différents contextes indique la localisation la plus probable connaissant les évènements précédents. La méthode utilise les définitions suivantes :

Définition 1 (observation) *Une observation o_{t_n} est une donnée structurée générée par un capteur ayant réagi à un événement e_{t_n} au temps t_n avec $n \in N$. À chaque observation o est associé le type de capteur $o.type$, l'origine du capteur $o.capteur$, et le rapport signal sur bruit (RSB) $o.rsb$ dans le cas de signaux numériques.*

Définition 2 (observations simultanées) *Deux observations a_{t_n} et b_{t_k} sont dites simultanées si $t_n = t_k$ et $a \neq b$.*

Définition 3 (hypothèse de localisation) *Étant donné un ensemble de localisations L défini par $L = \{Loc_1, \ldots, Loc_R\}$, et i, $i \in N$ tel que $h^i_{t_n} = Loc_i$, alors $h^i_{t_n}$ est le nœud hypothèse que l'habitant se trouve à la i^e localisation à l'instant t_n. Les hypothèses sont générées uniquement à partir des observations au temps t_n.*

Définition 4 (contexte d'occupation) *Étant donné un ensemble d'occupations P défini par $P = \{Piece_1, \ldots, Piece_R\}$, et i, $i \in N$ tel que $c^i = Piece_i$, alors c^i est le nœud de contexte d'occupation de la i^e pièce. La valeur d'activation du contexte c^i varie en fonction du temps et des hypothèses.*

Définition 5 (poids de relation) *$w \in [0,1]$ est la* force *de la relation entre deux nœuds du réseau. Ainsi, w_{o,h^i} est le poids du lien entre une observation o et l'hypothèse que l'habitant se trouve à la i^e localisation. w_{h^i,c^j} est le poids du lien entre l'hypothèse que l'habitant est à la i^e localisation et le contexte j.*

Définition 6 (fonction d'oubli) *La fonction d'oubli $f(t_n, t_{n-1}) = e^{-\frac{\Delta_t}{\tau}}$, avec $\Delta_t = t_n - t_{n-1}$, représente la décroissance d'un contexte au cours du temps. Elle permet de conserver une mémoire à court terme des contextes. Plus les observations seront espacées au cours du temps (p.ex., $\Delta_t > 3 \times \tau$), plus rapidement les contextes seront oubliés.*

5.2.1.1 Évolution du réseau dynamique

L'algorithme fonctionne ainsi :

1. pour toute nouvelle observation $o^k_{t_n}$, un nouveau nœud est créé ;
2. des nœuds hypothèses $h^i_{t_n}$ sont alors créés pour tout $i \in [1; r], r \leq R$ avant d'être connectés aux observations $o^k_{t_n}$ avec les poids w_{o^k,h^i} ;
3. les hypothèses $h^i_{t_n}$ sont connectées aux contextes de localisation c^j avec les poids w_{h^i,c^j} ;
4. les activations sont propagées des $o^k_{t_n}$ aux $h^i_{t_n}$ et l'activation des $h^i_{t_n}$ est calculée ;
5. les activations sont propagées de $h^i_{t_n}$ à c^j et l'activation de c^j est recalculée ;
6. le c^j avec la plus forte activation devient la localisation courante ;
7. tous les nœud $h^i_{t_n}$ et $o^k_{t_n}$ sont supprimés du réseau.

Un exemple de réseau dynamique est donné par la figure 5.1. À l'instant t_{n-2}, l'événement $e_{t_{n-2}}$ est capté par un capteur qui génère une observation $o_{t_{n-2}}$ dont 3 hypothèses sont déduites : $h^1_{t_{n-2}}$ avec une relation de 0,1 vers le contexte c^1, $h^2_{t_{n-2}}$, avec 0,6 vers c^2 et $h^3_{t_{n-2}}$ avec 0,3 vers c^3. S'il n'y a pas eu d'événement antérieur, c^2 est la localisation la plus probable. À l'instant t_{n-1}, l'événement $e_{t_{n-1}}$ est observé par deux capteurs. Les nœuds générés à t_{n-2} disparaissent, seul les contextes sont conservés. Si ces contextes sont encore actifs, leurs activations seront pondérées par $f(t_{n-1}, t_{n-2})$ auxquels on ajoute les activations générées par les nouvelles hypothèses. Dans le cas de c^2, l'activation sera incrémentée de $0,1 + 0,1 = 0,2$ car l'hypothèse $h^1_{t_{n-1}}$ est due à 2 événements simultanés $o^1_{t_{n-1}}$ et $o^2_{t_{n-1}}$. Dans le cas de c^4, l'hypothèse $h^2_{t_{n-1}}$ est due à une observation unique ; l'incrément sera égal au poids soit 0,8. La méthode s'appliquera ensuite de la même manière à l'instant t_n, etc.

5.2.1.2 Propagation de l'activation

L'activation recueillie par un nœud est classiquement définie (Crestani, 1997; Niessen et al., 2008) par la formule

$$n_i(t) = \sum_{i \neq j} w_{i,j} \times A^j(t) \tag{5.1}$$

où $w_{i,j}$ est le poids, j correspond à un voisin de i et $A^j(t)$ est l'activation de ce voisin au temps t. Bien entendu, un nœud qui a été activé par un voisin ne peut communiquer de nouveau cette activation à ce voisin. Dans notre cas, les activations sont toujours initiées par

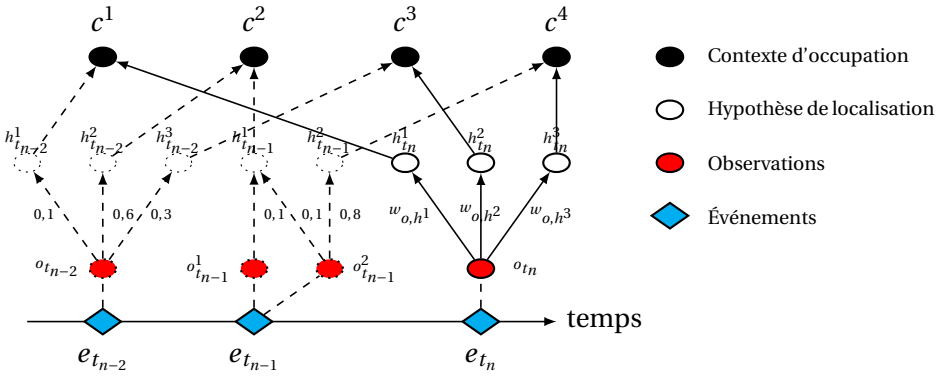

FIGURE 5.1 – Exemple de réseau dynamique

une observation o qui a à chaque fois une activation maximale $A^o(t) = 1$ quel que soit son type et se propage de bas en haut dans le réseau dynamique. Une fois l'activation $n(t)$ provenant des voisins accumulée, l'activation de sortie du nœud doit être calculée. Elle diffère selon le niveau des nœuds. Pour les hypothèses de localisation, l'activation $A^{h^i}(t) \in [0,1]$ est calculée selon la formule 5.2 :

$$A^{h^i}(t_n) = n_i(t_n) = \sum_{o \in O_{t_n}} w_{o,h^i} A^o(t_n) \tag{5.2}$$

où O_{t_n} est l'ensemble des observations simultanées à t_n liées à h^i, et $\sum_{o \in O_{t_n}} w_{o,h^i} \leq 1$. Pour les contextes d'occupation de pièce, l'activation de sortie résulte de l'accumulation des activations provenant des hypothèses et de l'activation précédente, pondérée par un facteur d'oubli. L'équation 5.3 décrit la loi d'activation du contexte de localisation A^{c^i} suite à une activation externe au temps t_n.

$$A^{c^i}(t_n) = n_i(t_n) \times [M - e^{-\frac{\Delta t}{\tau}} A^{c^i}(t_n - \Delta t)] + e^{-\frac{\Delta t}{\tau}} A^{c^i}(t_n - \Delta t) \tag{5.3}$$

où $A^{c^i}(t_n - \Delta t)$ est l'activation précédente, $M = 1$ est l'activation maximale, et $e^{-\frac{\Delta t}{\tau}}$ est la fonction d'oubli. Ainsi, si aucun événement n'apparait pendant $5 \cdot \tau$ secondes, l'activation des contextes pourra être considérée comme nulle. L'introduction de M permet de maintenir toutes les activations entre 0 et 1.

5.2.2 Calcul des relations entre les différentes couches du réseau

Le réseau dynamique étant constitué de deux niveaux bien définis (cf. figure 5.1), deux types de relations existent, la relation *observation-hypothèse* et la relation *hypothèse-contexte*. Les liens entres les différentes couches dépendent fortement de l'application et de l'environnement considéré, c'est pourquoi il est nécessaire d'introduire tout d'abord l'environnement applicatif pour faciliter la compréhension de ce chapitre.

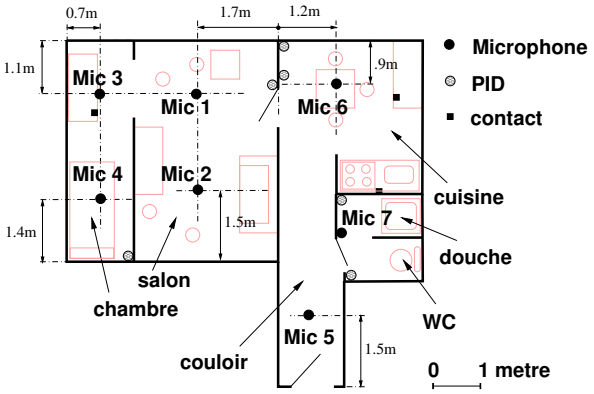

FIGURE 5.2 – Plan de l'appartement et position des capteurs.

5.2.2.1 Événements considérés dans l'environnement perceptif

Un exemple d'environnement perceptif considéré dans cette étude est l'Habitat Intelligent pour la Santé (HIS) de la Faculté de Médecine de Grenoble (Le Bellego et al., 2006). La figure 5.2 décrit l'appartement intelligent. Il est composé de 6 pièces contenant un ensemble de capteurs pour percevoir l'environnement : 7 microphones (Mic) placés dans le plafond (hauteur d'environ 3 mètres) ; 3 capteurs de contact sur des portes (CP) de meubles ; et 6 détecteurs de présence infrarouge (PID) placés sur les murs. Parmi les données recueillies par les capteurs (Mic, CP et PID), seuls les signaux sonores doivent subir un prétraitement. En effet, les données CP et PID fournissent une information booléenne déjà représentative d'événements. Les signaux audio ont été traités par le système AUDITHIS (Vacher et al., 2010). En bref, les événements audio sont détectés, en temps réel, et sont ensuite classés comme parole ou comme son de la vie courante (par exemple une chute d'objet). Les microphones étant omnidirectionnels, un événement sonore est généralement capté par plusieurs microphones en même temps. Pour chaque observation o de notre étude, on a ainsi : $o.type \in \{CP, Mic, PID\}$.

5.2.2.2 Relation hypothèse-contexte

La relation *hypothèse-contexte* est dans notre cas une relation univoque, car une hypothèse de localisation n'est liée qu'à une seule pièce. Par ailleurs, l'application visée étant la localisation d'un habitant dans une pièce, tous les nœuds contextes de pièces sont créés à l'initialisation du réseau et en font partie en permanence.

5.2.2.3 Relation observation-hypothèse

La relation *observation-hypothèse* est unidirectionnelle et de type 1-n (un vers plusieurs). Le poids et les hypothèses générés varient en fonction de la source des observations. Pour représenter cette relation dans le réseau, des informations de deux types sont considérées : celles qui sont dynamiques et dépendent des événements, et celles qui sont statiques car elles proviennent des connaissances *a priori* sur les capteurs.

Certaines connaissances sur l'environnement peuvent être utilisées pour définir les liens qui existent entre concepts. Pour une observation o, avec $o.type \in \{CP, PID\}$, un seul nœud hypothèse de localisation est créé. Les informations spatiales des PID et des CP sont univoques et certaines. Par exemple, l'ouverture du réfrigérateur ne peut se produire que si la personne se trouve dans la pièce. Par conséquent, un événement est lié avec un poids $w_{o,h} = 1$ à une hypothèse unique de localisation, et l'hypothèse est liée avec un poids de $w_{h,c} = 1$ au contexte. Il n'en va pas de même pour les informations des microphones. En effet, les microphones peuvent théoriquement capter toutes les ondes acoustiques générées dans les pièces d'une habitation. L'énergie du signal sonore détecté fournit bien une information dynamique sur la proximité de la source, mais il ne donne par contre aucune indication sur sa direction. Par exemple, un bruit capté à 1 mètre d'un microphone peut tout à fait avoir été généré dans une pièce différente de celle dans laquelle le microphone est placé. Pour prendre en compte ce phénomène, nous avons modélisé l'ambiguïté d'un microphone en fonction de sa distance avec les autres pièces par les poids w_{o,h^i} des relations dans le réseau dynamique. La valeur de ces poids a été calculée en estimant la probabilité $P(loc = i|Mic = j)$ que l'habitant soit à la i^e localisation sachant qu'un événement a été détecté sur le j^e microphone. Le poids de la relation entre l'hypothèse h^i et l'observation o est donné par la formule 5.4 :

$$w_{o^k,h^i}(t_n) = \frac{P(loc = i|o^k_{t_n}.capteur) \times o^k_{t_n}.rsb}{\sum_{l \in L} \sum_{o \in O_{t_n}} P(loc = l|o.capteur) \times o.rsb} \tag{5.4}$$

où O_{t_n} est l'ensemble des observations simultanées provenant des microphones à l'instant t_n, $P(loc|Mic)$ est une connaissance *a priori* décrite à la section 5.2.2.4, L est l'ensemble des localisations possibles, et $o.rsb$ est le rapport signal sur bruit linéarisé de o. Ainsi, w_{o^k,h^i} est un poids normalisé ($\sum_{i,k} w_{o^k,h^i} = 1$) qui représente le lien entre les observations sonores simultanées et les hypothèses de localisation.

5.2.2.4 Acquisition des informations statiques

Pour calculer la probabilité $P(loc = i|Mic = j)$, deux approches ont été testées, l'approche naïve et l'approche par analyse statistique de corpus. Pour l'approche naïve, la référence sonore a été choisie à 1 mètre au dessous du microphone. À partir de cette référence, un cercle de 2 mètres de rayon a été tracé autour de chaque microphone, ce qui correspond à une atténuation de $-6dB$ en considérant l'atténuation quadratique classique. Au delà de ce cercle, la perte d'énergie est supérieure à 75%. Le poids est calculé selon la surface de l'intersection entre le cercle et les pièces, avec une pénalité de 2 lorsqu'une cloison est traversée.

L'autre approche est statistique ; elle calcule directement les probabilités à partir du corpus annoté. Le tableau 5.1 indique les poids obtenus pour l'approche naïve et pour l'approche statistique.

Le mélange d'informations *a priori* et d'informations dynamiques permet une meilleure désambiguïsation. Par exemple, dans le cas de deux observations sonores simultanées captées par le micro 6 de la cuisine et le micro 7 de la salle de bains avec un RSB identique égal

à $12dB$, les formules 5.2 et 5.4 et l'a priori naïf donnent pour l'activation

$$A^{h^{cuisine}} = rsb \times (.87 + .5)/[rsb \times (.87 + .5 + .18 + .13 + .03 + .18 + .1)] = 0,69 \quad (5.5)$$

qui sera supérieure à celle de la salle de bains $A^{h^{sdb}} = 0,09$ bien que le rapport signal sur bruit soit le même.

5.3 Résultats

L'approche a été testée sur des données réelles acquises lors d'expériences (Fleury et al., 2010b) destinées à évaluer la reconnaissance automatique des activités de la vie quotidienne d'une personne à son domicile dans le but de pouvoir détecter automatiquement une perte d'autonomie. Dans le HIS (cf. section 5.2.2.1), 15 volontaires ont effectué chacun au moins une fois 7 activités prédéfinies (entre autres, dormir, faire sa toilette, communiquer avec l'extérieur) sur une période d'environ 1 heure, sans consigne sur la manière ou l'ordre dans lequel ces activités devaient être effectuées. Les données ont ensuite été annotées grâce à des caméras placées uniquement à cet effet.

Pour chaque enregistrement, les observations enregistrées par les capteurs CP, PID et Mic ont activé un réseau dynamique. Les traces d'activation de chaque contexte ont été conservées, et l'emplacement de départ de l'expérimentation est supposé connu (ici la cuisine).

Les performances de localisation ont été évaluées à l'échelle de la seconde. À chaque seconde, le contexte de localisation de plus haut poids est comparé à la vérité « terrain ». S'il y a correspondance, alors il s'agit d'un vrai positif (VP), sinon il s'agit d'une confusion. Cela nous permet de générer une table de confusion à partir de laquelle le taux d'exactitude est calculé par $TE = \frac{nb(VP)}{nb(test)}$, où $nb(test)$ correspond à la durée de l'évaluation exprimée en secondes.

Le tableau 5.2 montre les résultats sur les 47226 secondes (13h 7min 6s) d'enregistrement des 15 participants, pour chaque combinaison de capteurs. Dans le cas sans *a priori* (c.-à-d., $P(loc = i|Mic = j) = 1$ si le capteur j est dans i^e pièce 0 sinon), l'ajout des contacts de porte CP améliore légèrement la précision obtenue avec seulement les capteurs à infrarouge PID, mais l'utilisation de l'information sonore dégrade légèrement les performances (88,2% au lieu de 88,9%). Lorsqu'un *a priori* sur les relations événement sonore - hypothèse de localisation est ajouté et calculé par la méthode naïve, alors, comme dans la méthode statistique,

	$P(Loc\|Mic)$ estimé par la méthode naïve							$P(Loc\|Mic)$ estimé à partir du corpus						
Mic	1	2	3	4	5	6	7	1	2	3	4	5	6	7
chambre	0,14	0,07	0,70	0,85				0,28	0,29	0,42	0,43	0,25	0,18	0,20
salon	0,86	0,93	0,27	0,14	0,01	0,13	0,03	0,59	0,56	0,47	0,41	0,07	0,07	0,09
cuisine			0,03	0,02	0,10	0,87	0,50	0,05	0,08	0,06	0,09	0,45	0,63	0,37
douche					0,06		0,18	0,06	0,05	0,04	0,04	0,09	0,04	0,10
WC					0,06		0,18	0,01	0,01	0,01	0,01	0,12	0,05	0,21
couloir					0,77		0,10		0,02	0,01	0,02	0,03	0,03	0,02

TABLE 5.1 – Estimation de $P(Loc|Mic)$

Capteurs	Mic	PID	CP	Mic+CP	PID+Mic	PID+CP	PID+Mic+CP
sans *a priori*	25,7	**88,9**	26,5	32,8	87,7	89,4	88,2
a priori naïf	30,2	88,9	26,5	34,1	89,0	89,4	**89,5**
a priori naïf sujet 4	25,3	92,8	19,5	25,4	96,8	92,8	**96,8**
a priori statistique	30,9	88,9	26,5	34,8	89,7	89,4	**90,1**

TABLE 5.2 – Exactitude (%) avec plusieurs combinaisons de sources

nous observons une amélioration des performances. La connaissance *a priori* du lien entre un son et un capteur permet de corriger un grand nombre d'erreurs de localisation. Le gain reste cependant faible, lorsque le changement de pièce n'est pas détecté par les PID mais que les paroles sont bien identifiées, l'information sonore est utile pour améliorer la localisation (comme dans le cas du sujet 4) en compensant le défaut de sensibilité des infrarouges. Cela illustre l'intérêt de la combinaison de plusieurs sources. Pour le participant 4 et dans le cas *a priori* naïf, la valeur de l'activation des contextes de localisation pour chacune des occurrences d'événements est tracée sur la sous-figure (a) de la figure 5.3 pour les 60 premiers événements. La sous-figure (b) montre, sur le même intervalle, la localisation réelle de la personne (trait plein) et la localisation estimée (trait plein surmonté d'un point).

Dans cet extrait, le participant est dans le séjour (événements 1 à 9), puis il se rend à la cuisine (événements 10 à 15) avant de revenir au séjour. Les événements donnant des informations sur cette transition sont sonores (moins précis que les PID). La méthode estime la localisation dans la douche (événements 16 à 18) avant de trouver la vraie localisation. Ensuite, le participant reste dans le salon (événements 16 à 60) pour effectuer un appel téléphonique. Au cours de la période dans le séjour, les deux contextes *Séjour* et *Chambre* s'affrontent, car la conversation téléphonique se tient à proximité de la limite du salon et de la chambre à coucher. Le participant étant resté presque immobile pendant la conversation, il n'y a pas eu d'événement détecté par les PID et les microphones ont donc constitué la seule source d'information.

5.4 Discussion

Ce chapitre a présenté une méthode de fusion d'informations multisource multimodale et temporelles par propagation d'activation dans un réseau dynamique qui permet de modéliser la connaissance statique (liée à l'organisation de l'espace) et dynamique (temporelle). Cette approche permet de fusionner des sources non visuelles avec une grande souplesse et n'impose pas une perception continue de l'espace ni de patron d'événements. De plus, cette méthode permet de prendre en compte facilement un *a priori* dans le réseau afin de tenir compte de connaissances expertes.

Les résultats montrent que la fusion d'informations par propagation d'activation multisource présente un intérêt, même dans le cas où les sources ont une très bonne précision, comme c'est le cas des capteurs à infrarouge (mais avec une sensibilité imparfaite). L'utilisation de sources de localisation plus incertaines, telles que la détection de parole, peut alors

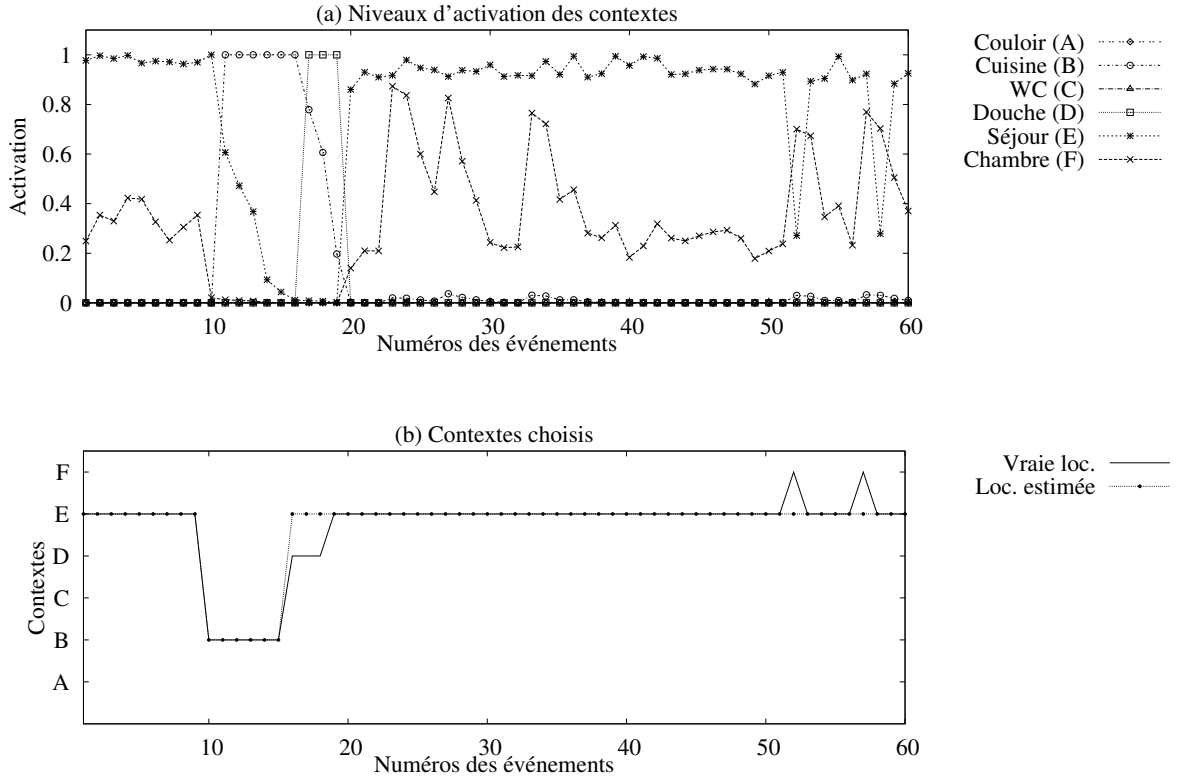

FIGURE 5.3 – Extrait de l'évolution des niveaux d'activation des contextes (a) et des contextes choisis pour les 6 localisations possibles (b).

améliorer les performances. Malgré l'ambiguïté inhérente aux microphones omnidirectionnels, la prise en compte de la connaissance a priori du lien entre un son et un microphone permet de corriger un grand nombre d'erreurs de localisation.

Cependant, l'utilisation unique des informations de parole conduit cette source à être très peu sensible. Dans notre étude, quasiment 60% des confusions sont dues à des périodes durant lesquelles le sujet ne parle pas, n'utilise pas d'objet et durant lesquelles ses mouvements ne sont pas captés par les PID. Nous prévoyons donc d'inclure la classification de différents types de sons afin de capter les mouvements générateurs de bruits qui ne peuvent pas être détectés par les PID.

Plusieurs autres pistes d'amélioration de cette méthode vont être suivies dans l'avenir. Une première piste consiste à agir sur la constante de temps en fonction de la sémantique des événements et des lieux. Par exemple, une détection dans le couloir devrait avoir une constante de temps plus faible que dans le cas des toilettes. À cet effet, nous prévoyons d'extraire les constantes de temps potentielles à partir des données. Sur le plus long terme, la méthode présentée pourra être améliorée en utilisant la théorie de Dempster-Shafer. Une fusion d'informations serait alors possible pour déterminer la localisation par consensus entre le niveau de croyance des sources d'information et les poids d'activation des hypothèses. Cette théorie de combinaison de preuves offre un cadre de travail plus formel que la combinaison de poids. Enfin, nous prévoyons d'appliquer cette méthode à la classification des sons de la vie courante en utilisant le contexte de la localisation pour désambiguïser les hypothèses de classification des sons.

CHAPITRE 6

Classification des activités de la vie quotidienne à partir de données multimodales en utilisant une méthode de type SVM

La connaissance des activités de la vie quotidienne peut être utile à deux niveaux. Comme cela a déjà été présenté à la section 1.3, la connaissance de l'évolution à long terme des AVQ peut être utile d'une part pour aider à déterminer le niveau d'autonomie d'une personne, et, d'autre part pour détecter certains troubles lorsque la fréquence de certaines activités devient anormale (Virone et al., 2003). Par ailleurs, la connaissance de l'activité que la personne est en train de réaliser ou des dernières activités qu'elle a réalisées peut être, au même titre que la localisation, une information intéressante et utile pour un système de prise de décision en domotique.

Ce chapitre présente les aspects principaux d'une expérimentation qui a été réalisée dans le HIS du laboratoire TIMC-IMAG décrit à la figure 2.2 de la section 2.3.1. Cette expérimentation ainsi que ses résultats sont décrits plus précisément dans l'article (Fleury et al., 2010a) qui est reproduit dans l'annexe A.3.

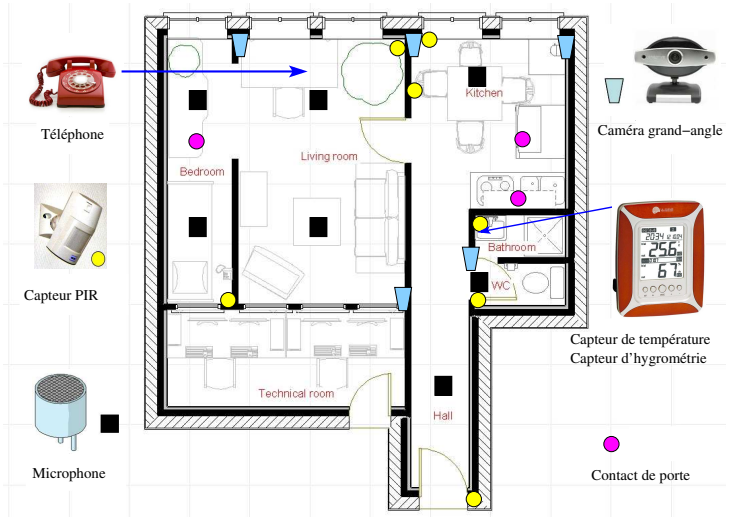

FIGURE 6.1 – Position des capteurs dans l'appartement d'étude de la Faculté de Médecine

Modalité	Paramètre sélectionné	Information apportée
Actimètre	Temps passé en pourcentage dans des postures variées (debout, assis, couché) ou à marcher	Dormir (être couché), se reposer (être couché ou assis), prendre un repas (marcher puis s'asseoir)
Microphone	Nombre d'événements par classe de sons (parole, cris, ouverture et fermeture de portes, vaisselle, pas, sonnerie de téléphone, chute d'objets, bris de verre), et nombre d'événements par microphone	Communication (sonnerie de téléphone et parole), préparation des repas (vaisselle), repos (sons de TV ou de radio dans le salon)
Détecteur de présence à infrarouge (PIR)	Temps passé en pourcentage dans chaque pièce (mobilité) et nombre d'événements par détecteur (agitation)	Dormir et se vêtir/se dévêtir (dans la chambre, immobile dans le 1er cas et mobile dans le second), cuisiner (dans la cuisine), faire sa toilette (dans la salle de bains), éliminer (dans les toilettes), se reposer ou communiquer (dans le salon)
Contact de porte	Temps ouvert en pourcentage et position prédominante (ouvert ou fermé) dans la trame temporelle	Cuisiner (utilisation des placards et du frigo), se vêtir/se dévêtir (utilisation du tiroir de la commode)
Capteur environnemental	Mesure différentielle de la température et de l'hygrométrie toutes les 15 minutes	Faire sa toilette (utilisation du radiateur ou de la douche)

TABLE 6.1 – Modalités des différents type de capteur

6.1 Paramètres utilisés

Les informations étaient fournies par les différents capteurs situés dans l'appartement : des capteurs de présence infrarouge (PIR) présents dans chaque pièce, des contacts placés sur certaines portes, la commode et les placards, un capteur environnemental dans la salle de bains, et 7 microphones placés au niveau du plafond. La personne volontaire pour l'expérience portait de plus un actimètre fixé sur un vêtement. La figure 6.1 montre l'emplacement précis des différents capteurs utilisés. Les caméras ne sont utilisées que pour l'annotation du corpus enregistré ; il n'y en avait aucune dans la salle de bains et les toilettes.

La classification des activités ne peut se faire sur les données brutes extraites des capteurs car elles seraient trop différentes d'une expérimentation à l'autre. Il faut donc définir des paramètres pour lesquels cette variabilité est moins grande. Il a donc été nécessaire de réaliser des expérimentations préalables de courte durée qui ont permis de calculer un grand nombre de paramètres avant de sélectionner ceux qui ont paru les plus adaptés. Les paramètres retenus pour les différents capteurs ainsi que les informations apportées sont présentées dans le tableau 6.1.

Par ailleurs, la taille de la trame temporelle utilisée pour extraire le paramètres a été fixée à 3 minutes, ce qui correspond à la durée minimale estimée d'une activité. Chaque trame est décrite par un vecteur $X = [\alpha_{1_1}, \alpha_{1_2}, \alpha_{1_3}, \ldots, \alpha_{5_1}, \alpha_{5_2}]$ de 42 paramètres α_{i_j} où $i \in 1\ldots5$ désigne chacune des différentes modalités, et où j est un index désignant chacun des paramètres à l'intérieur de la modalité. Les données recueillies sont très hétérogènes ; elles ont donc été normalisées pour pouvoir être utilisées avec un algorithme SVM. Le jeu de données normalisé ainsi obtenu a une moyenne nulle et un écart type unitaire.

6.2 Protocole expérimental et données recueillies

Chacune des personnes participant à l'expérience a effectué 7 AVQ qui doivent être réalisées dans la vie courante au moins une fois par jour :

1. *Dormir* : un lit était disponible dans la chambre.
2. *Préparer et prendre le petit déjeuner* : le matériel et la nourriture nécessaires étaient disponibles dans la cuisine. Chacun pouvait procéder à sa convenance et devait ensuite nettoyer la cuisine et la vaisselle.
3. *Se vêtir et se dévêtir* : des vêtements à enfiler étaient disponibles dans la chambre.
4. *Se reposer* : la personne pouvait à sa guise lire un livre, un journal, écouter la radio, regarder la télé, etc.
5. *Faire sa toilette* : la personne devait au moins se laver les mains et les dents, la porte devait être fermée.
6. *Élimination* : pour cette activité, la personne utilisait les toilettes.
7. *Communication* : cette activité consistait à recevoir un appel téléphonique et à avoir une conversation. Dans le protocole, la personne était appelée 5 fois et devait répondre en lisant à chaque fois une des conversations qui lui avait été fournie.

Chaque activité devait être réalisée au moins une fois dans l'ordre souhaité par la personne, et sans impératif de durée. Une visite préalable du site d'expérimentation permettait à la personne de prendre connaissance des lieux et de vérifier qu'elle pouvait trouver tout ce qui était nécessaire à la réalisation des activités.

Treize personnes jeunes et en bonne santé ont participé à l'expérimentation, 6 femmes et 7 hommes. La moyenne d'âge était de 30,4 ans (24-43 min-max), la taille moyenne 1,76 m (1,62-1,92 min-max), et, le poids moyen 69 kg (57-80 min-max). La durée moyenne de l'expérimentation par participant était de 51 min 40 s (23 min 11 s-1h 35 min 44 s min-max). Le tableau 6.2 montre la répartition des trames entre les différentes activités.

Activité	Nombre de trames	Pourcentage (%)
Dormir	49	21,1
Préparer et prendre le petit déjeuner	45	19,4
Se vêtir et se dévêtir	16	6,9
Se reposer	73	31,4
Faire sa toilette	14	6
Élimination	16	6,9
Communication	19	8,1
Total	**232**	**100**

TABLE 6.2 – Distribution des trames suivant les différents AVQ dans le corpus

La répartition n'est pas bien équilibrée entre les différentes activités ; cela est essentiellement dû au protocole expérimental qui a laissé une grande liberté aux participants. Les activités qui ont réclamé le plus de temps ont été : *Dormir*, *Préparer et prendre le petit déjeuner* et *Se reposer*. Le temps passé à faire sa toilette a été en général très bref. Étant donné

que l'ordre entre activités n'était pas imposé, certains se sont couchés seulement quelques minutes avant de se lever pour commencer une nouvelle activité. Si une personne a passé entre 2 et 3 minutes pour réaliser une activité, la fenêtre a été considérée comme une fenêtre de pleine activité de 3 minutes.

6.3 Méthode de classification SVM

6.3.1 SVM multiclasses

La méthode de classification SVM (*Support Vector Machine*, ou plus exactement Séparateur à Vaste Marge) retenue dite « un contre un » a consisté à construire un classifieur pour toutes les combinaisons de paires (C_i, C_j) des N classes à différencier, avec $0 < i \leq N$ et $0 < j < i$ (Knerr et al., 1990). Avec cette méthode, seule une part de l'espace relativement réduite demeure indifférenciée ; un vote majoritaire permet ensuite de déterminer la classe correspondant à un nouveau point : $C = max_{k=1...N} Card(\{y_{i,j}\} \cap \{k\})$ où $y_{i,j}$ est la décision obtenue pour ce point par le SVM entraîné pour distinguer entre les classe i et j. En cas d'égalité, la classe retenue est celle procurant la marge minimale.

Une méthode introduite par Guermeur (2002) consiste à résoudre un problème d'optimisation unique, c'est à dire à créer une réelle séparation multiclasse avec un modèle de régression linéaire multivarié (jeu d'hyperplans de cardinalité égale au nombre de classes). Cette méthode est plus complexe et gourmande en temps calcul du fait de la détermination de séparations par résolution de problèmes linéaires. Une comparaison de méthodes a montré par ailleurs la similarité des résultats entre la méthode multiclasse et la méthode « un contre un » (Hsu and Lin, 2002).

6.3.2 Séparation linéaire

Construire un SVM revient à construire un hyperplan séparateur d'équation $\vec{w}^T \vec{x} + w_0 = 0$, où $\vec{w}^T$ et w_0 sont les paramètres à calculer pour définir cet hyperplan. À partir de cet hyperplan, nous construisons la fonction $f()$ qui permet de définir à laquelle des 2 classes appartient un vecteur $\vec{y}$:

$$f(\vec{y}) = \begin{cases} +1 & \text{si} \quad \vec{w}^T \vec{y} + w_0 \geq 0 \\ -1 & \text{si} \quad \vec{w}^T \vec{y} + w_0 < 0 \end{cases} \tag{6.1}$$

Étant donné un ensemble des exemples $\{(\vec{x}_1, e_1), \ldots, (\vec{x}_N, e_N)\}$ constituant le jeux d'apprentissage, avec $\vec{x}_i$ les vecteurs de caractéristiques de dimension D, et $e_i \in \{-1, +1\}$ les étiquettes (si l'étiquette vaut $+1$, alors le vecteur appartient à la classe à reconnaître), l'hyperplan choisi doit maximiser la distance entre les points les plus proches de chaque classe tout en restant un séparateur. Cela revient à minimiser $\frac{1}{2}\|\vec{w}\|^2$ sous les contraintes $e_i(\vec{w}^T \vec{x}_i + w_0) \geq 1$, pour $i \in 1 \ldots N$. Cela peut se résoudre par la méthode classique des multiplicateurs de La-

grange, où le lagrangien est donné par :

$$L(\vec{w}, w_0, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{k=1}^{N} \alpha_k \left[e_i(\vec{w}^T \vec{x}_i + w_0) - 1 \right] \tag{6.2}$$

où les coefficients α_k sont les multiplicateurs de Lagrange. Le lagrangien doit être minimisé par rapport à $\vec{w}$ et w_0, et maximisé par rapport aux coefficients α_k. La résolution du problème montre que l'hyperplan optimal ne dépend que des n_s vecteurs supports $\vec{x}_k$ du problème :

$$\vec{w} = \sum_{k=1}^{n_s} \alpha_k \, e_k \, \vec{x}_k \tag{6.3}$$

En notant $\langle ., . \rangle$ le produit scalaire de deux vecteurs, le système d'équations 6.1 devient :

$$f(\vec{y}) = signe\left(\vec{w}^T \vec{y} + w_0\right) = signe\left(\sum_{k=1}^{n_s} e_k \, \alpha_k \, \langle \vec{x}_k, \vec{y} \rangle + w_0\right) \tag{6.4}$$

6.3.3 Cas général : séparation non linéaire

Dans la plupart des cas, il n'est pas possible de séparer des classes par un hyperplan ; on doit alors appliquer une fonction non linéaire ϕ aux vecteurs d'entrée $\vec{x} \in X$ qui ont ainsi une image $\phi(\vec{x})$ dans un nouvel espace $\phi(X)$. Le problème consiste alors à construire dans ce nouvel espace l'hyperplan de séparation défini par : $\vec{w}^T \phi(\vec{x}) + w_0 = 0$ où $\vec{w}^T$ et w_0 sont les paramètres définissant cet hyperplan. Dans les équations 6.1 et 6.4 qui permettent respectivement la classification d'un nouveau point $\vec{x}$ et la résolution du problème, le vecteur $\vec{x}$ apparait toujours à l'intérieur d'un produit scalaire.

En pratique, on ne connait pas la fonction ϕ, aussi Aizerman et al. (1964) ont décrit une famille de fonctions « noyau » K ayant les propriétés d'un produit scalaire dans un espace de grande dimension $K(\vec{x}i, \vec{x}j) = \langle \vec{x}_i, \vec{x}_j \rangle$. Cette fonction doit être choisie *a priori* car il n'existe pas à l'heure actuelle de méthode permettant de la déterminer. La détermination de l'hyperplan est alors obtenue en remplaçant le produit scalaire dans les équation 6.2 et 6.4 par la fonction noyau $K(.,.)$. Le théorème de Mercer (Mercer, 1909) explicite les conditions que K doit satisfaire pour être une fonction noyau : elle doit être symétrique, semi-définie positive. L'exemple le plus simple de fonction noyau est le noyau linéaire pour lequel on n'opère pas de changement d'espace : $K(\vec{x}_i, \vec{x}_j) = \vec{x}_i^T . \vec{x}_j$. Les fonctions noyau utilisées de manière classique les plus courantes sont :

1. le noyau polynomial : $K(\vec{x}_i, \vec{x}_j) = \left(\vec{x}_i^T . \vec{x}j + 1\right)^p$
2. le noyau gaussien : $K(\vec{x}_i, \vec{x}_j) = \exp\left(\frac{-\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma^2}\right)$

6.4 Résultats de classification

La méthode a été testée en utilisant un noyau polynomial de degré 2 et un noyau gaussien avec un protocole « tous sauf un » (*leave one out*) étant donné la faible taille du jeu de

données. En ce qui concerne le noyau gaussien, les valeurs des paramètres ont été choisies de façon à minimiser le taux d'erreur global. Les résultats, en termes de taux de bonne classification, correspondant à ces 2 méthodes sont donnés, dans le tableau 6.3.

Activité	Noyau polynomial (%)	Noyau gaussien (%)
Dormir	77,6	93,9
Préparer et prendre le petit déjeuner	84,4	97,8
Se vêtir et se dévêtir	56,2	75
Se reposer	76,7	78,1
Faire sa toilette	50	64,3
Élimination	68,7	93,75
Communication	89,5	89,3
Total	**75,9**	**86,2**

TABLE 6.3 – Taux de bonne classification pour les noyaux polynomial et gaussien

Les résultats sont similaires pour les deux noyaux. Les 2 classes *Se vêtir et se dévêtir* et *Faire sa toilette* sont nettement moins bien classifiées que les autres classes, ce qui semble dû à leur faible représentation dans le jeu de données. Quand les vecteurs support sont enlevés de la base d'apprentissage, la construction du modèle est altérée et cela prend d'autant plus d'importance que la classe est représentée par peu de données. Nous obtenons 75,9% de trames bien classifiées pour le noyau polynomial et 86,2% pour le noyau gaussien.

La matrice de confusion donnée dans le tableau 6.4 montre en effet que les erreurs pour la classe *Se vêtir et se dévêtir* sont dues à une redistribution sur les classes *Dormir* et *Se reposer* qui sont les classes les plus représentées. Cette matrice montre aussi que les classes *Faire sa toilette* et *Éliminer* sont très proches et peuvent facilement être confondues. Cela est dû en partie au fait qu'elles se situent obligatoirement dans 2 pièces petites et contiguës et que l'activité *Éliminer* suppose en principe tout de suite après une activité *Faire sa toilette* pour se laver les mains.

Les erreurs de l'activité *Se reposer* sont en partie dues à une confusion avec la classe *Dormir*, car dans les 2 cas il n'y a pas ou peu d'indication des capteurs sensibles au mouvement (actimètre, PIR) ou aux sons, seule la présence dans la pièce occupée est connue, et il y a donc très peu d'informations redondantes. Donc, si la dernière indication d'un capteur est erronée, rien ne permettra d'apporter une correction.

Pour terminer, en ce qui concerne l'activité *Communiquer*, les confusions ont lieu avec *Se vêtir/se dévêtir* (qui se peut se passer dans la même pièce), *Préparer et prendre le petit*

		Taux de bonne classification						
		Dormir	Se reposer	Se vêtir/ se dévêtir	Se nourrir	Éliminer	Faire sa toilette	Communiquer
Activité	Dormir	**93,9%**	0%	0%	0%	0%	0%	0%
	se reposer	13,8%	**78,1%**	1,3%	1,3%	4,2%	1,3%	0%
	Se vêtir/se dévêtir	6,25%	12,5%	**75%**	0%	0%	0%	6.25%
	Se nourrir	0%	0%	2,2%	**97.8%**	0%	0%	0%
	Éliminer	0%	0%	0%	6,25%	**93,75%**	0%	0%
	Faire sa toilette	7,1%	0%	0%	7,1%	21,5%	**64,3%**	0%
	Communiquer	0%	5,3%	5,9%	0%	0%	0%	**89,5%**

TABLE 6.4 – Matrice de confusion pour le noyau gaussien avec une optimisation de σ

déjeuner (confusion possible dans la discrimination entre parole et son de vaisselle), et *Se reposer* car les fauteuils et le téléphone sont dans la même pièce.

En résumé, le pourcentage global de trames bien classifiées était de 86%, et les classes les mieux classifiées sont celles qui ont le plus grand nombre de représentants et sont les mieux représentées dans le corpus de test.

6.5 Discussion

Cette expérimentation a permis de montrer qu'il était possible de classifier parmi 7 AVQ prédéfinies les activités de la vie quotidienne de personnes agissant dans un appartement de manière réaliste. Le taux de bonnes classifications est de 75% avec un noyau polynomial et de 86% avec un noyau gaussien.

Les résultats sont différents selon les classes, les classes les mieux représentées conduisant aux meilleurs résultats. Ces résultats sont proches des meilleurs résultats publiés qui sont de 88%, mais qui utilisent beaucoup plus de capteurs et marquent tous les objets (Philipose et al., 2004). La méthode que nous avons présentée est plus proche de celle développée par Kröse et al. (2008) bien que cette dernière ne considère que deux activités.

L'utilisation d'autres paramètres pouvant être plus pertinents est à envisager, et doit permettre d'améliorer la robustesse de la classification. Le choix des paramètres adéquats peut être fait à partir des scores obtenus, par exemple par des méthodes de sélection d'attributs (Vacher et al., 2010). Des paramètres apportant une information temporelle pourraient s'avérer très utiles, comme par exemple le temps de présence dans la pièce où se trouve la personne. Il semble par ailleurs difficile d'apporter une connaissance *a priori* au classifieur SVM (Lauer and Bloch, 2008) sans risquer de réduire la capacité de généralisation de l'algorithme et de fausser la détection d'activités lorsque elles ont eu lieu, par exemple, à des moments ou dans des lieux qui ne correspondent pas à une activité normale.

Une amélioration de la méthode pourrait consister à introduire une classe de rejet qui ne correspondrait à aucune activité particulière, et qui pourrait permettre une détection d'activité de type AVQ ainsi qu'une classification guidée par les activités. Un découpage en trames de 1 minute seulement devrait alors mieux convenir. Par ailleurs, il pourrait s'avérer pertinent d'utiliser comme entrée supplémentaire la localisation de la personne obtenue par une méthode différente prenant en compte les aspects temporels. Ce pourrait être le cas de la méthode des réseaux dynamiques temporels présentée au chapitre 5.

CHAPITRE 7

Quelques perspectives de recherche et projets à venir

Mes perspectives de recherche s'inscrivent dans la continuité des thèmes présentés dans les chapitres précédents en rapport avec l'analyse sonore et multimodale appliquée à l'assistance à domicile. Deux projets financés par l'ANR vont me permettre de continuer à développer ces activités en ce qui concerne la reconnaissance de la parole et des sons, et l'analyse multimodale.

7.1 Projets ANR

7.1.1 SWEET-HOME

Le projet SWEET-HOME que j'ai conçu et dont je suis le responsable scientifique est financé par l'ANR (programme VERSO) et a débuté en 2010. Le partenariat associe :

1. 2 partenaires académiques : le LIG au travers des équipes GETALP et MULTICOM, et l'ESIGETEL au travers de l'équipe ANASON ;
2. 3 partenaires industriels de type PME : THEORIS, Technosens et Caméra-contact.

Ce projet vise à développer un nouveau système pour le bâtiment intelligent. Ce système doit s'appuyer d'une part sur les protocoles de communication standards du domaine (réseau KNX), et, d'autre part sur l'analyse des informations sonores dans l'appartement. Ce projet vise à réaliser trois objectifs qui devront permettre de mettre une personne seule à son domicile en mesure de piloter son environnement :

- en procurant une interaction homme-machine naturelle (commandes vocales et tactiles) ;
- en facilitant l'inclusion sociale grâce à des dispositifs spécialisés développés par les partenaires industriels ;
- en procurant un sentiment de sécurité par détection des situations de détresse.

L'étude d'acceptabilité présentée au chapitre 3 a été réalisée dans ce cadre et a montré l'intérêt de cette approche. Outre cette étude d'usage, la contribution du LIG porte plus particulièrement sur la reconnaissance et l'extraction des ordres vocaux, la classification des sons, et l'interprétation de la situation suivie de la prise de décision. Dans le cadre de ce projet, j'ai recruté un post-doctorant, Benjamin Lecouteux tout récemment recruté comme MDC à l'IUT-2 de Grenoble, et, je coencadre un étudiant en thèse, Pedro Chahuara. Ces travaux ont déjà fait l'objet de publications sur la reconnaissance de la parole en conditions

distantes (Lecouteux et al., 2011a) et la localisation à partir d'informations multimodales et multisource (Chahuara et al., 2010).

7.1.2 CIRDO

Le projet CIRDO porté par le LIRIS et dont je suis le responsable en ce qui concerne la participation du LIG est financé par l'ANR (programme TECSAN) et a débuté au début de l'année 2011. Ce projet consolide une collaboration déjà existante entre le LIG et le LIRIS. Le partenariat associe :

1. 3 partenaires académiques : le LIRIS (équipe SAARA), le LIG (équipe GETALP), et le GRePS ;
2. 3 partenaires industriels du domaine de l'assistance à domicile : CATEL, Technosens et ISARP.

Ce projet vise à renforcer le lien social des personnes âgées isolées à domicile ou vivant de manière autonome dans des institutions avec leur entourage proche : famille, amis, et, personnel aidant. La contribution du LIG portera sur deux aspects. Le premier aspect concernera l'étude des caractéristiques particulières de la parole émise par les personnes âgées et les conséquences qui en découlent sur la reconnaissance automatique de ce type de parole. Le deuxième aspect concerne la reconnaissance de scènes à partir d'informations sonores et visuelles. Dans le cadre de ce projet, je coencadre un étudiant en thèse, Frédéric Aman, qui a pour tâche d'étudier les caractéristiques spécifiques de la parole des personnes âgées et l'adaptation des systèmes de reconnaissance automatique à ce type de population. Les sons et les paroles spontanées (onomatopées, jurons, etc.) seront pris en compte afin de traiter les évènements non langagiers.

7.2 Analyse multimodale

7.2.1 Reconnaissance des AVQ

Une approche permettant de classifier les activités de la vie quotidienne a été présentée au chapitre 6. Étant donné l'intérêt que peut présenter la connaissance de l'évolution du comportement d'une personne au travers des AVQ, aussi bien pour apprécier une baisse du taux d'autonomie que pour aider à détecter certaines pathologies, la détection et la reconnaissance automatiques des AVQ serait un apport positif.

Je compte poursuivre dans cette voie grâce à une collaboration que je suis en train de monter avec une équipe de recherche de l'École des Mines de Douai. Les corpus disponibles et enregistrés lors des différentes phases du projet SWEET-HOME pourront être mis à profit pour cette recherche.

7.2.2 Prise de décision pour l'assistance à domicile

Un module de décision doit générer des actions en fonction du contexte, par exemple des actions domotiques dans le cas du système SWEET-HOME. Par exemple, les lumières doivent être éteintes en cas d'absence, et la lumière du couloir doit être progressive la nuit. La prise de décision contextuelle a fait l'objet des travaux de Portet et al. (2005) et Callens et al. (2008) dans le domaine du pilotage en ligne d'algorithmes de traitement du signal en utilisant des arbres de décision. Une piste pourrait consister à explorer les théories et méthodes d'aide à la décision adaptées au traitement d'informations temporelles de différentes natures, imprécises et incertaines, telles que la théorie de Dempster-Shafer (Shafer, 1976). L'intérêt d'une approche telle que celle de Dempster-Shafer pour ce problème de décision est la possibilité de baser une décision sur une combinaison de preuves (Hong et al., 2009). Cette approche formalisée présenterait l'avantage de permettre l'insertion d'*a priori* dans les modèles.

7.2.3 Reconnaissance de scènes

La détection des situations anormales nécessitant de générer une alerte est un des services essentiels d'un système de surveillance à domicile. Ces situations peuvent être détectées en analysant les primitives générées par les étages de traitement audio et vidéo. Par exemple, la détection de cris et de bruits sourds (chute d'un objet lourd) dans un intervalle de temps réduit permet d'inférer l'occurrence d'une chute. Bien entendu, comme dans tout système de suivi, ce genre d'inférence ne doit pas générer de fausses alarmes (Chambrin, 2001).

Ce problème, très courant dans les cas où l'information acquise est incomplète et/ou incertaine, peut être surmonté en utilisant des méthodes de fusion de données (ou analyse multimodale) qui tireraient partie de l'information vidéo et audio ainsi que des informations sur le patient. Ces informations sont connues pour être complémentaires. En effet, la vidéo apporte des informations sur le suivi, la position et les mouvements de la personnes ; ces informations sont nécessairement partielles à cause des angles morts introduits par le champ de vision de la caméra. Le son apporte des informations langagières, les limitations sont essentiellement dues à la distance entre la personne et le microphone.

Ce thème de recherche sera abordé dans le cadre de la collaboration avec le Professeur S. Bouakaz et l'équipe SARAA du LIRIS, dans le cadre du projet CIRDO.

7.3 Analyse des sons et de la parole

7.3.1 Reconnaissance de la parole en conditions distantes

L'analyse de la parole enregistrée en condition distante est encore loin d'être résolue, même si certains travaux ont pu conduire à des avancées, notamment, comme cela a été présenté au chapitre 4, en tirant profit d'enregistrements simultanés obtenus par plusieurs

microphones. En conditions réelles, la parole est souvent mélangée avec d'autres sons parasites selon un rapport variable. Ces sons parasites peuvent être dus à une source connue et stable comme la radio, mais ce n'est pas le cas le plus fréquent (bruit de la rue, etc.).

Une première possibilité de solution consisterait donc à faire appel à des techniques inspirées des méthodes de séparation de sources aveugles ou de l'analyse en composantes indépendantes (ICA). D'autres solutions pourraient consister à utiliser des modèles acoustiques appris en conditions bruitées ou partiellement bruitées et ensuite d'utiliser des méthodes s'appuyant sur l'analyse factorielle (*factor analysis*) (Matrouf et al., 2007).

7.3.2 Classification des sons

La classification des sons utilisant des méthodes purement statistiques a montré ses limites lorsqu'on l'utilise en conditions réelles, ce qui est dû à la très grande diversité des sons rencontrés par rapport à ce qui se passe pour la parole qui correspond à une source d'émission de signal unique, même si on observe malgré tout une dispersion entre locuteurs suivant l'âge, la morphologie et le sexe de chacun.

La diversité et la très grande variabilité des sources de sons de la vie courante rendent par ailleurs difficile la prise en compte de l'évolution temporelle du signal, comme cela est fait dans le cas de la parole : évolution temporelle au niveau du phonème modélisé par des HMM, et ensuite regroupement de phonèmes modélisés par le dictionnaire phonétique et les modèles de langage.

Des méthodes de classification hiérarchique prenant en compte certains paramètres peu sensibles au bruit comme la forme du signal temporel, la fréquence fondamentale, et la durée du signal, devraient permettre d'obtenir une classification plus robuste. Un classifieur basé sur la prise en compte de la morphologies des sons (Schaeffer, 1966) a déjà été expérimenté lors du projet *Sample Orchestrator* à l'IRCAM pour la description, l'indexation et la classification automatique des contenus sonores et musicaux dans des bases de données pour le cinéma (Vinet, 2009; Peeters and Deruty, 2010). Les corpus de sons de la vie courante enregistrés en conditions réalistes dans le cadre du projet SWEET-HOME pourront servir de base à cette recherche.

7.3.3 Reconnaissance de la parole des personnes âgées

Les changements de la voix en fonction de l'âge ont fait l'objet de plusieurs études, mais le comportement des systèmes de reconnaissance de la parole a été peu étudié sauf par Vipperla et al. (2010) dans le cas de la langue anglaise, qui ont disposé des enregistrements de plaidoiries prononcées à la Cour Suprême des États-Unis par des procureurs identiques sur une décennie, et a ainsi pu évaluer les évolutions de performances en fonction de l'âge pour un même locuteur ainsi que l'efficacité de l'adaptation au locuteur. Une autre étude dans le cas de la langue anglaise a montré qu'il n'y a pas uniquement une évolution au niveau acoustique, mais aussi au niveau du vocabulaire (Vipperla et al., 2009). Par exemple, l'augmentation des mots hors vocabulaire provoque une baisse importante du taux de reconnaissance.

La reconnaissance automatique de la parole a aussi été étudiée dans le cas du japonais (Baba et al., 2004).

La parole émise par le locuteur âgé se caractérise principalement par une modification de la fréquence fondamentale, un manque d'énergie, une instabilité dans la production de certaines consonnes, et une augmentation du bruit (Gorham-Rowan and Laures-Gore, 2006; Hooper and Craidis, 2009). Il sera donc intéressant, dans le cadre du projet CIRDO, d'approfondir ces études dans le cas du français. Cela nécessitera l'acquisition préalable d'un corpus adapté auprès de personnes âgées, corpus qui devra comporter d'une part une partie de texte lu utilisable pour l'apprentissage de modèles, et d'autre part de la conversation spontanée qui pourra être utilisée par exemple pour l'analyse de scènes, et ensuite la prise de décision de mise en relation avec un proche ou de génération de compte rendu d'activité. Cela sera utile en outre pour déterminer quelles sont les phrases significatives d'un appel à l'aide ou d'une détresse qu'il conviendra de mettre en évidence pour déclencher une mise en relation ou une intervention.

Par ailleurs, l'enregistrement de sons non langagiers dans la conversation spontanée pourra aussi contribuer à la connaissance des signes émis volontairement ou non par la personne et susceptibles de véhiculer une information concernant une situation anormale (chute, immobilité prolongée, ...), une anxiété, une souffrance, etc.

7.3.4 Évolution de la richesse du vocabulaire

Nous observons un vieillissement de la population française et des pays développés. La maladie d'Alzheimer et les maladies apparentées progressent inexorablement avec l'âge : à partir de 85 ans, une femme sur quatre et un homme sur cinq sont touchés. Un des symptômes de la maladie est la réduction progressive du vocabulaire utilisé par la personne.

Les systèmes de reconnaissance de la parole pourraient être utilisés pour aider à déterminer l'évolution de la gravité de la maladie en mesurant l'évolution de la taille du vocabulaire utilisé par la personne lors d'activités contraintes mais distractives, comme par exemple des jeux. Cela pourrait tirer profit des enregistrements et études faites dans le cadre du projet CIRDO sur la parole des personnes âgées.

7.3.5 Cas des personnes âgées qui reviennent à la langue de leur enfance

Les personnes très âgées reviennent facilement à la langue de leur petite enfance. L'équipe GETALP a acquis des compétences en systèmes de reconnaissance automatique de la parole multilingue. Même si c'est à une échéance plus lointaine, ces compétences pourraient utilement être mises à profit pour aider le personnel soignant confronté à ce genre de problèmes.

Notons qu'il existe déjà aux USA un système de traduction de l'oral, destiné au personnel de santé anglophone voulant communiquer avec des patients hispanophones et leurs proches, il s'agit de Converser for HealthCare , de Spoken Translation Inc., créé par M. Seligman qui a participé au projet CSTAR.

Cette étude pourrait être faite dans le cadre d'une collaboration franco-japonaise suite à des contacts récents avec le professeur Kawahara de Kyoto.

Bibliographie

Aizerman, M., Braverman, E., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25 :821–837.

Aldrich, F. (2003). Smart homes : Past, present and future. In Harper, R., editor, *Inside the Smart Home*, pages 17–39. Springer London.

Amoretti, M., Copelli, G., Matrella, G., Grossi, F., and Baratta, S. (2010). The PERSONA AAL Platform : Deployment in the italian pilot site of Bardi. In *AALIANCE Conference on Ambient Assisted Living : Technology and Innovation for Ageing Well*, Malaga, Spain.

Anastasiou, A., Boulos, M. K., and consortium, C. (2008). Barriers to the adoption of telehealth and the complete ambient assisted living experiment (CAALYX). In *RAATE 2008—Recent Advances in Assistive Technology & Engineering*, Coventry, UK.

Aswath, D., D'cunha, J., Ahmed, S. T., and Davulcu, H. (2005). Boosting item keyword search with spreading activation. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI'05*, pages 704–707.

Augusto, J. C. (2009). *Agents and Artificial Intelligence*, volume 67, Part1 of *Communication in Computer and Information Science*, chapter Past, Present and Future of Ambient Intelligence and Smart Environments, pages 3–15. Springer.

Baba, A., Lee, A., Saruwatari, H., and Shikano, K. (2002). Speech recognition by reverberation adapted acoustic models. In *ASJ General Meeting*, pages 27–28.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., and Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan*, 87, Part 2(7) :49–57.

Badii, A. and Boudy, J. (2009). CompanionAble - integrated cognitive assistive & domotic companion robotic systems for ability & security. In *Proceedings of the First Congress of the Société Française des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG'09)*, pages 18–20, Troyes.

Bahadori, S., Cesta, A., Grisetti, G., Iocchi, L., Leone, R., Nardi, D., Oddi, A., Pecora, F., and Rasconi, R. (2004). RoboCare : Pervasive Intelligence for the Domestic Care of the Elderly. *Intelligenza Artificiale*, 1(1) :16–21.

Beaudin, J., Intille, S., and Morris, M. (2006). To track or not to track : User reactions to concepts in longitudinal health monitoring. *Journal of Medical Internet Research*, 8(4) :29–47.

Beritelli, F., Casale, S., Ruggeri, G., and Serrano, S. (2002). Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors. *IEEE Signal Processing Letters*, 9(3) :85–88.

Berndt, D. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases (KDD'94)*, pages 359–370.

Bian, X., Abowd, G. D., and Rehg, J. M. (2005). Using sound source localization in a home environment. In *Proceedings of Pervasive 2005, International conference*, pages 19–36, Munich, Germany.

Bonhomme, S., Campo, E., Estève, D., and Guennec, J. (2008). Prosafe-extended, a telemedicine platform to contribute to medical diagnosis. *Journal of Telemedicine and Telecare*, 14(3) :116–119.

Brdiczka, O., Langet, M., Maisonnasse, J., and Crowley, J. (2009). Detecting human behavior models from multimodal observation in a smart home. *IEEE Transactions on Automation Science and Engineering*, 6(4) :588 –597.

Callejas, Z. and López-Cózar, R. (2009). Designing smart home interfaces for the elderly. *SIGACCESS Newsletter*, 95.

Callens, L., Carrault, G., Cordier, M.-O., Fromont, E., Portet, F., and Quiniou, R. (2008). Intelligent adaptive monitoring for cardiac surveillance. In *Proceeding of the 2008 conference on ECAI 2008 : 18th European Conference on Artificial Intelligence*, pages 653–657. IOS Press.

Caon, D., Amehraye, A., Razik, J., Chollet, G., Mokbel, C., and Andreao, R. (2010). Experiments on Acoustic Model supervised adaptation and evaluation by K-Fold Cross Validation technique. In *Proceedings of the 5th International Symposium on I/V Communications and Mobile Networks*, Rabat, Maroc.

Chahuara, P., Vacher, M., and Portet, F. (2010). Localisation d'habitant dans un environnement perceptif non visuel par propagation d'activation multisource. In *Proceedings of MAJECSTIC*, pages 1–8, Bordeaux, France.

Chambrin, M.-C. (2001). Alarms in the intensive care unit : how can the number of false alarms be reduced ? *Critical Care*, 5(4) :184–188.

Chan, M., Campo, E., Estève, D., and Fourniols, J.-Y. (2009). Smart homes — current features and future perspectives. *Maturitas*, 64(2) :90–97.

Chan, M., Estève, D., Escriba, C., and Campo, E. (2008). A review of smart homes- present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91(1) :55–81.

Chen, J., Kam, A. H., Zhang, J., Liu, N., and Shue, L. (2005). Bathroom activity monitoring based on sound. In *Proceedings of Pervasive 2005, International conference*, pages 47–61, Munich, Germany.

Cook, D. J., Youngblood, M., III, E. O. H., Gopalratnam, K., Rao, S., Litvin, A., and Khawaja, F. (2003). Mavhome : an agent-based smart home. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications*, page 521, Los Alamitos, CA, USA.

Cortes, U., Barrue, C., Martinez, A., Urdiales, C., Campana, F., Annicchiarico, R., and Caltagirone, C. (2010). Assistive technologies for the new generation of senior citizens : the share-it approach. *International Journal of Computers in Healthcare*, 1(1) :35–65.

Cowling, M. (2004). *Non-Speech Environmental Sound Classification System for Autonomous Surveillance*. PhD thesis, Griffith University.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6) :453–482.

Dalal, S., Alwan, M., Seifrafi, R., Kell, S., and Brown, D. (2005). A rule-based approach to the analysis of elders activity data : Detection of health and possible emergency conditions. In *AAAI Fall 2005 Symposium*.

Demiris, G., Rantz, M., Aud, M., Marek, K., Tyrer, H., Skubic, M., and Hussam, A. (2004). Older adults' attitudes towards and perceptions of "smart home" technologies : a pilot study. *Medical Informatics and the Internet in Medicine*, 29(2) :87–94.

Deng, L., Acero, A., Plumpe, M., and Huang, X. (2000). Large-vocabulary speech recognition under adverse acoustic environments. In *Proceedings of ICSLP-2000*, volume 3, pages 806–809, Beijing, China. ISCA.

Dufaux, A. (2001). *Detection and Recognition of Impulsive Sound Signals*. PhD thesis, Faculté des Sciences de l'Université de Neuchâtel, Suisse.

Dugheanu, R. (2011). Evaluation des outils pour la reconnaissance automatique de la parole adaptée aux personnes âgées. Master's thesis, Master professionnel des Sciences du Langage, Université Stendhal, Grenoble 3.

Duong, T., Phung, D., Bui, H., and Venkatesh, S. (2009). Efficient duration and hierarchical modeling for human activity recognition. *Artificial Intelligence*, 173(7-8) :830–856.

Edwards, W. and Grinter, R. (2001). At home with ubiquitous computing : Seven challenges. In *Proceedings of Ubicomp 2001 : Ubiquitous Computing*, pages 256–272, Atlanta, Georgia, USA.

Fiscus, J.-G. (1997). A post-processing system to yield reduced word error rates : Recognizer Output Voting Error Reduction (ROVER). In *Proceedings of IEEE Workshop ASRU*, pages 347–354.

Fleury, A., Vacher, M., and Noury, N. (2010a). SVM-based multimodal classification of activities of daily living in health smart homes : sensors, algorithms and first experimental results. *Information Technology in Biomedicine, IEEE Transactions on*, 14(2) :274 –283.

Fleury, A., Vacher, M., Portet, F., Chahuara, P., and Noury, N. (2010b). A multimodal corpus recorded in a health smart home. In *Proceedings of Multimodal Corpora and Evaluation, LREC MMC Workshop*, pages 99–105, Malta.

Friedland, G., Yeo, C., and Hung, H. (2010). Dialocalizaton : Acoustic speaker diarization and visual localization as joint optimization problem. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(4) :27 :1–27 :18.

Fugger, E., Prazak, B., Hanke, S., and Wassertheurer, S. (2007). Requirements and ethical issues for sensor-augmented environments in elderly care. In *Proceedings of the 4th International Conference on Universal Access in Human-Computer-Interaction*, pages 887–893.

Gerosa, M., Giuliani, D., and Brugnara, F. (2009). Towards age-independent acoustic modeling. *Speech Communication*, 51(6) :499–509.

Giroux, S., Pigot, H., Mayers, A., Lefebvre, B., Rialle, V., and Noury, N. (2002). Smart house for frail and cognitive impaired elders. In *UbiCog '02 : First International Workshop on Ubiquitous Computing for Cognitive Aids*, Göteborg, Sweden.

Gorham-Rowan, M. and Laures-Gore, J. (2006). Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of Communication Disorders*, 39 :171–184.

Goumopoulos, C., Kameas, A., Hagras, H., Callagan, V., Gardner, M., Minker, W., Weber, M., Bellik, Y., and Meliones, A. (2008). ATRACO : Adaptive and Trusted Ambient Ecologies. In *Proceedings of the 2nd IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO), Workshop on Pervasive Adaptation (PERADA)*.

Guermeur, Y. (2002). Combining discriminant models with new multiclass SVMs. *Pattern Analysis and Applications*, 5(2) :168–179.

Haigh, K. Z. and Yanco, H. (2002). Automation as caregiver : a survey of issues and technologies. In *Proceedings of the AAAI-02 Workshop Automation as Caregiver : The Role of Intelligent Technology in Elder Care*, pages 39–53.

Hamill, M., Young, V., Boger, J., and Mihailidis, A. (2009). Development of an automated speech recognition interface for personal emergency response system. *Journal of NeuroEngineering and Rehabilitation*, 26(6).

Hong, X., Nugent, C., Mulvenna, M., McClean, S., Scotney, B., and Devlin, S. (2009). Evidential fusion of sensor data for activity recognition in smart homes. *Pervasive and Mobile Computing*, 5(3) :236–252.

Hooper, C. R. and Craidis, A. (2009). Normal Changes in the Speech of Older Adults : You've still got what it takes ; it just takes a little longer ! *Perspectives on Gerontology*, 14.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transaction on Neural Networks*, 13 :415–425.

Ibartz, A., Bauer, G., Casas, R., Marco, A., and Lukowicz, P. (2008). Design and evaluation of a sound-based water flow measurement system. In *Proceedings of the 3rd European Conference on Smart Sensing and Context (LNCS 5279)*, pages 41–54.

Ibarz, A., Casas, R., Marco, A., Garrido, Y., Falco, J., and Roy, A. (2007). Health and social considerations in ambient intelligence design. In *Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence*, pages 91–96, Zaragoza, Spain.

Intille, S. S. (2002). Designing a home of the future. *IEEE Pervasive Computing*, 1(2) :76–82.

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

Jaimes, A., Sebe, N., and Gatica-Perez, D. (2006). Human-centered computing : a multimedia perspective. In *Proceedings of the 14th annual ACM international conference on Multimedia, MULTIMEDIA'06*, pages 855–864, New York, NY, USA. ACM.

Jambon, F. (2009). User evaluation of mobile devices : In-situ versus laboratory experiments. *International Journal of Mobile Computer-Human Interaction*, 1(2) :56–71.

Jianmin Jiang, A. G. and Zhang, S. (2009). HERMES : a FP7 funded project towards computer-aided memory management via intelligent computations. In *Proceedings of the 3rd Symposium of Ubiquitous Computing and Ambient Intelligence*, pages 249–253.

Kang, M.-S., Kim, K. M., and Kim, H.-C. (2006). A questionnaire study for the design of smart homes for the elderly. In *Proceedings of Healthcom 2006*, pages 265–268.

Katz, S. and Akpom, C. (1976). A measure of primary sociobiological functions. *International Journal of Health Services*, 6(3) :493–508.

Kent, S. I., Larson, K., and Tapia, E. M. (2003). Designing and evaluating technology for independent aging in the home. In *International Conference on Aging, Disability and Independence*.

Klein, M., Schmidt, A., and Lauer, R. (2007). Ontology-centred design of an ambient middleware for assisted living : the case of SOPRANO. In *Proceedings of 30th Annual German Conference on Artificial Intelligence (KI 2007), Towards Ambient Intelligence : Methods for Cooperating Ensembles in Ubiquitous Environments*, Osnabrück, Germany.

Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited : a stepwise procedure for building and training a neural network. In *Neurocomputing : Algorithms, Architectures and Applications*. Springer-Verlag.

Kortenkamp, D. and Chown, E. (1993). A directional spreading activation network for mobile robot navigation. In *Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, pages 218–224.

Koskela, T. and Väänänen-Vainio-Mattila, K. (2004). Evolution towards smart home environments : empirical evaluation of three user interfaces. *Personal and Ubiquitous Computing*, 8 :234–240.

Kratt, J., Metze, F., Stiefelhagen, R., and Waibel, A. (2004). *DAGM-Symposium'04*, chapter Large vocabulary audio-visual speech recognition using the Janus speech recognition toolkit, pages 488–495. Springer Berlin.

Kröse, B., van Kasteren, T., Gibson, C., and van den Dool, T. (2008). CARE : Context Awareness in Residences for Elderly. In *The 6th International Conference of the International Society of Gerontology*, Pisa, Italy.

Lacombe, A., Rocaries, F., Dietrich, C., Baldinger, J., Boudy, J., Delavault, F., Descatha, A., Baer, M., and Ozguler, A. (2005). Open technical platform prototype and validation process model for patient at home medical monitoring system. In *BioMedsim*, Linköping, Sweden.

Lauer, F. and Bloch, G. (2008). Incorporating prior knowledge in support vector machines for classification : A review. *Neurocomputing*, 71 :1578–1594.

Lawton, M. and Brody, E. (1969). Assesment of older people : self-maintaining and instrumental activities of daily living. *Gerontologist*, 9 :179–186.

Le Bellego, G., Noury, N., Virone, G., Mousseau, M., and Demongeot, J. (2006). A model for the measurement of patient activity in a hospital suite. *IEEE Transactions on Information Technology in Biomedicine*, 10(1) :92 – 99.

Lecouteux, B., Linarès, G., Bonastre, J., and Nocéra, P. (2006). Imperfect transcript driven-speech recognition. In *Proceedings of InterSpeech'06*, pages 1626–1629, Pittsburg, Pennsylvania, USA.

Lecouteux, B., Linarès, G., Estève, Y., and Gravier, G. (2008). Generalized driven decoding for speech recognition system combination. In *Proceedings of IEEE ICASSP 2008*, pages 1549–1552, Las Vegas, Nevada, USA.

Lecouteux, B., Vacher, M., and Portet, F. (2011a). Distant speech recognition for home automation : preliminary experimental results in a smart home. In *IEEE SPED 2011*, pages 41–50, Braşov, Romania.

Lecouteux, B., Vacher, M., and Portet, F. (2011b). Distant speech recognition in a smart home : comparison of several multisource ASRs in realistic conditions. In *Proceedings of Interspeech 2011*, pages 1–4, Florence, Italy.

Lesser, V., Atighetchi, M., Benyo, B., Horling, B., Raja, A., Vincent, R., Wagner, T., Xuan, P., and Zhang, S. X. (1999). The UMASS intelligent home project. In *Proceedings of the Third International Conference on Autonomous Agents*, pages 291–298, Seattle, USA.

Li, S. and Wrede, B. (2007). Why and how to model multi-modal interaction for a mobile robot companion. In *AAAI Spring Symposium 2007 on Interaction Challenges for Intelligent Assistants*.

Linarès, G., Nocéra, P., Massonié, D., and Matrouf, D. (2007). The LIA speech recognition system : from 10xRT to 1xRT. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue, TSD'07*, pages 302–308, Pilsen, Czech Republic.

Lines, L. and Hone, K. (2006). Multiple voices, multiple choices : older adult's evaluation of speech output to support independent living. *Gerontechnology Journal*, 2(5) :78–91.

Litvak, D., Zigel, Y., and Gannot, I. (2008). Fall detection of elderly through floor vibrations and sound. In *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, pages 4632–4635.

López-Cózar, R. and Callejas, Z. (2010). *Handbook for Ambient Intelligence and Smart Environments*, chapter Multimodal dialogue for ambient intelligence and smart environments, pages 559–579. Springer US.

Lutolf, R. (1992). Smart home concept and the integration of energy meters into a home-based system. In *Proceedings of the International Conference on Metering Apparatus and Tariffs for Electricity Supply*, pages 277–278.

Maier, E. and Kempter, G. (2010). ALADIN – a magic lamp for the elderly ? In *Handbook of Ambient Intelligence and Smart Environments*, pages 1201–1227.

Matrouf, D., Scheffer, N., Fauve, B., and Bonastre, J.-F. (2007). A straightforward and efficient implementation of the factor analysis model for speaker verification. In *Proceedings of Interspeech 2007*, pages 1242–1245, Antwerp, Belgium.

Medjahed, H., Istrate, D., Boudy, J., and Dorizzi, B. (2009). A fuzzy logic system for home elderly people monitoring (EMUTEM). In *Proceedings of the 10th WSEAS international conference on Fuzzy systems*, pages 69–75, Prag, Czech Republic. World Scientific and Engineering Academy and Society (WSEAS).

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A*, 209 :415–446.

Mozer, M. C. (2005). *Smart environments : Technologies, protocols, and applications*. J. Wiley & Sons.

Mäyrä, F., Soronen, A., Vanhala, J., Mikkonen, J., Zakrzewski, M., Koskinen, I., and Kuusela, K. (2006). Probing a proactive home : Challenges in researching and designing everyday smart environments. *Human Technology*, 2 :158–186.

Nain, G., Barais, O., Fleurquin, R., and Jézéquel, J.-M. (2009). Entimid : un middleware aux services de la maison. In *3ème Conférence Francophone sur les Architectures Logicielles (CAL'O9)*, Nancy, France.

Nemer, E., Goubran, R., and Mahmoud, S. (2001). Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*, 9(3).

Niessen, M. E., van Maanen, L., and Andringa, T. C. (2008). Disambiguating sounds through context. In *Proceedings of the second IEEE International Conference on Semantic Computing, ICSC2008*, pages 88–95. IEEE Computer Society.

Panek, P., Edelmayer, G., Oliver, D., Maguire, M., McCrindle, R., Nissen, J., Nussbaum, G., Stanek, O., Victor, C., and Zagler, W. (2007). ENABLE - a wrist-worn device with integrated accessible services to support old people living independently and safely at home. In *Proceedings of the 9th Europ. Conf. for the Advancement of Assistive Technology in Europe (AAATE)*, pages 758–762, San Sebastian, Spain.

Patterson, D., Etzioni, O., Fox, D., and Kautz, H. (2002). Intelligent ubiquitous computing to support Alzheimer's patients : enabling the cognitively disabled. In *UbiCog '02 : First International Workshop on Ubiquitous Computing for Cognitive Aids*, Göteborg, Sweden.

Peeters, G. and Deruty, E. (2010). Sound indexing using morphological description. *IEEE Transactions on Audio Speech and Language Processing*, 18(3) :675–687.

Philipose, M., Fishkin, K. P., Perkowitz, M., Patterson, D. J., Fox, D., Kautz, H., and Hahnel, D. (2004). Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 3(4) :50–57.

Popescu, M., Li, Y., Skubic, M., and Rantz, M. (2008). An acoustic fall detector system that uses sound height information to reduce the false alarm rate. In *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, pages 4628–4631.

Portet, F., Ignacio Hernandez, A., and Carrault, G. (2005). Evaluation of real-time QRS detection algorithms in variable contexts. *Medical & Biological Engineering & Computing*, 43(3) :381–387.

Portet, F., Vacher, M., Golanski, C., Roux, C., and Meillon, B. (2011). Design and evaluation of a smart home voice interface for the elderly - acceptability and objection aspects. *Personal and Ubiquitous Computing*, pages 1–30. (in press).

Rantz, M., Porter, R., Cheshier, D., Otto, D., Servey, C., Johnson, R., Aud, M., Skubic, M., Tyrer, H., He, Z., Demiris, G., Alexander, G., and Taylor, G. (2008). TigerPlace, a State-Academic-Private project to revolutionize traditional long-term care. *Journal of Housing For the Elderly*, 22(1) :66–85.

Rialle, V. (2007a). *Technologie et Alzheimer : appréciation de la faisabilité de la mise en place de technologies innovantes pour assister les aidants familiaux et pallier les pathologies de type Alzheimer*. PhD thesis, Université René Descartes-Paris 5.

Rialle, V. (2007b). Technologies nouvelles susceptibles d'améliorer les pratiques gérontologiques et la vie quotidienne des malades âgés et de leur famille. Technical report, Ministère de la Santé et des Solidarités, 74 pages.

Rialle, V., Ollivet, C., Guigui, C., and Hervé, C. (2008). What do family caregivers of Alzheimer's disease patients desire in smart home technologies ? Contrasted results of a wide survey. *Methods of Information in Medicine*, 47(1) :63–69.

Rougui, J., Istrate, D., and Souidene, W. (2009). Audio sound event identification for distress situations and context awareness. In *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '09)*, pages 3501–3504, Minneapolis, USA.

Rumeau, P., Rialle, V., and Noury, N. (2006). A priori evaluation of acceptance of an activity monitoring device for the disabled elderly people using the HIS as a model. In *Proceedings of the 4th International Conference on Smart Homes and Health Telematics, ICOST2006, Smart Homes and Beyond*, pages 130–137, Belfast, Ireland. IOS Press.

Schaeffer, P. (1966). *Traité des objets musicaux*. Le Seuil.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

Skubic, M., Alexander, G., Popescu, M., Rantz, M., and Keller, J. (2009). A smart home application to eldercare : current status and lessons learned. *Technology and Health Care*, 17(3) :183–201.

Soler, V., Penalver, A., Zuffanelli, S., Roig, J., and Aguiló, J. (2010). Domotic hardware infrastructure in PERSONA project. In *International Symposium on Ambient Intelligence (ISAmI 2010)*.

Sweet-Home (2010). Le projet SWEET-HOME : Système Domotique d'Assistance au Domicile. http://sweet-home.imag.fr/.

Tang, P. and Venables, T. (2000). Smart homes and telecare for independent living. *Journal of Telemedicine and Telecare*, 6(1) :8–14.

Tapia, E. M., Intille, S. S., and Larson, K. (2004). Activity recognition in the home using simple and ubiquitous sensors. *Pervasive Computing*, 2 :158–175.

Tran, H.-D. and Li, H. (2009). Sound event classification based on feature integration, recursive feature elimination and structured classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 177–180, Taipei, Taiwan.

Vacher, M., Fleury, A., Guirand, N., Serignat, J.-f., and Noury, N. (2009). Speech recognition in a smart home : some experiments for telemonitoring. In Corneliu Burileanu, H.-N. T., editor, *From Speech Processing to Spoken Language Technology*, pages 171–179, Constanta (Romania). Publishing House of the Romanian Academy.

Vacher, M., Fleury, A., Portet, F., Serignat, J.-F., and Noury, N. (2010). *New Developments in Biomedical Engineering*, chapter Complete Sound and Speech Recognition System for Health Smart Homes : Application to the Recognition of Activities of Daily Living, pages 645 – 673. Intech Book.

Vacher, M., Fleury, A., Serignat, J.-F., Noury, N., and Glasson, H. (2008). Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment. In *Proceedings of Interspeech 2008*, pages 496–499, Brisbane, Australia.

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, M., Portet, F., Fleury, A., and 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.

Vacher, M., Serignat, J., Chaillol, S., Istrate, D., and Popescu, V. (2006). *Lecture Notes in Artificial Intelligence, 4188/2006*, chapter Speech and Sound Use in a Remote Monitoring System for Health Care, pages 711–718. Springer Berlin/Heidelberg.

Vacher, M., Serignat, J.-F., and Chaillol, S. (2007). Sound classification in a smart room environment : an approach using GMM and HMM methods. In C. Burileanu, H.-N. T., editor, *Advances in Spoken Language Technology*, pages 135–146, Iasi, Romania.

Valin, J.-M. (2007). On adjusting the learning rate in frequency domain echo cancellation with double-talk. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3) :1030–1034.

Van-Thinh, V., Bremond, F., and Thonnat, M. (2003). Automatic video interpretation : a novel algorithm for temporal scenario recognition. In *Proceedings of the 18th international joint conference on Artificial intelligence*, pages 1295–1300, Acapulco, Mexico. Morgan Kaufmann Publishers Inc.

Vaufreydaz, D., Bergamini, C., Serignat, J.-F., Besacier, L., and Akbar, M. (2000). A new methodology for speech corpora definition from Internet documents. In *Proc. LREC'2000, 2nd Int. Conf. on Language Ressources and Evaluation*, pages 423–426, Athens, Greece.

Vergados, D., Alevizos, A., Mariolis, A., and Caragiozidis, M. (2008). Intelligent services for assisting independent living of elderly people at home. In *Proceedings of the 1st International Conference on Pervasive Technologies Related to Assistive Environments*, pages 704–707.

Vinet, H. (2009). Le projet Sample Orchestrator. http ://anasynth.ircam.fr/home/projects/sample-orchestrator. Projet ANR RIAM (2006-2009).

Vipperla, R., Renals, S., and Frankel, J. (2008). Longitudinal study of ASR performances on ageing voices. In *Proceedings of INTERSPEECH*, pages 2550–2553, Brisbane, Australia.

Vipperla, R., Renals, S., and Frankel, J. (2010). Ageing voices : the effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, Article ID 525783 :1–10.

Vipperla, R., Wolters, M., Georgila, K., and Renals, S. (2009). Speech input from older users in smart environments : challenges and perspectives. In *Proceedings of the 5th International on ConferenceUniversal Access in Human-Computer Interaction. Part II : Intelligent and Ubiquitous Interaction Environments*, UAHCI '09, pages 117–126, Berlin, Heidelberg. Springer-Verlag.

Virone, G., Istrate, D., Vacher, M., Serignat, J., Noury, N., and Demongeot, J. (2003). First steps in data fusion between a multichannel audio acquisition and an information system for home healthcare. In *Proceedings of the International Conference of the IEEE Engineering in Medecine and Biology Society*, pages 1364–1367, Cancun, Mexico.

Weiser, M. (1991). The computer for the 21st century. *Scientific American*, 265(3) :66–75.

Wilpon, J. and Jacobsen, C. (1996). A study of speech recognition for children and the elderly. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 349–352.

Wölfel, M. and McDonough, J. (2009). *Distant Speech Recognition*. John Wiley and Sons, 573 pages.

Xi, C., Zhihai, H., Keller, J. M., Anderson, D., and Skubic, M. (2006). Adaptive silhouette extraction in dynamic environments using fuzzy logic. In *Proceedings of the International Conference on Fuzzy Systems*, pages 236–243, Vancouver, Canada.

Zouba, N., Bremond, F., Thonnat, M., Anfosso, A., Pascual, E., Mallea, P., Mailland, V., and Guerin, O. (2009). A computer system to monitor older adults at home : preliminary results. *Gerontechnology Journal*, 8(3) :129–139.

Index

ANNEXE A

Publications

A.1 Conférence EUSIPCO 2004

"Sound Detection and Classification through Transient Models using Wavelet Coefficient Trees"

M. VACHER, D. ISTRATE AND J.-F. SERIGNAT

12th European Signal Processing Conference EUSIPCO, Vienne (Autriche), pp. 1171-1174, 6-10 septembre 2004.

SOUND DETECTION AND CLASSIFICATION THROUGH TRANSIENT MODELS USING WAVELET COEFFICIENT TREES

Michel Vacher, Dan Istrate and Jean-François Serignat

CLIPS - IMAG (UMR CNRS-INPG-UJF 5524, Team GEOD
385, rue de la Bibliothèque - BP 53, 38041 Grenoble cedex 9, France (Europe)
phone: +33 4 7663 5795, fax: +33 4 7663 5552, email: Michel.Vacher@imag.fr
web: www-clips.imag.fr

ABSTRACT

Medical Telesurvey needs human operator assistance by smart information systems. Usual sound classification may be applied to medical monitoring by use of microphones in patient's habitation. Detection is the first step of our sound analysis system and is necessary to extract the significant sounds before initiating the classification step. This paper proposes a detection method using transient models, based upon dyadic trees of wavelet coefficients to insure short detection delay. The classification stage uses a Gaussian Mixture Model classifier with classical acoustical parameters like MFCC. Detection and classification stages are evaluated in experimental recorded noise condition which is non-stationary and more aggressive than simulated white noise and fits with our application. Wavelet filtering methods are proposed to enhance performances in low signal to noise ratios.

1. INTRODUCTION

In this paper a sound detection/classification method is presented. This method has been developed as part of a medical telesurvey system intended for home hospitalization. The aim of this system is to detect a distress situation of the patient using sound analysis. In case of distress a medical center is automatically called with the aim in view to give assistance to the patient. The decision of call is taken by a data fusion system from smart sensors and particularly a sound system as explained in [1].

Each sound produced in the apartment is characteristic of:

- a patient's activity: the patient is locking the door, or he is walking in the bedroom,
- the patient's physiology: he his having a cough,
- a possible distress situation for the patient: a scream or a glass breaking are suddenly appearing.

If the system has a good ability of classification for such sounds, it will be feasible to know if the patient is needing help. Several usual sound classes needed for this application have been defined and a corpus has been recorded in our laboratory.

Before sound classification, it is necessary in a first step to establish the start and the stop time of the sound to classify in the environmental noise. The precision of this 2 times must be sufficient to allow the classification step good performances. In the context of audio signal encoding the input signal can be decomposed into "tonal", "transient" and "stochastic" components as described by Daudet in [2][5]; our problem is restricted to transient detection for which large wavelet coefficients are more easily interpreted as transients.

Proposed methods are based on trees of wavelet coefficients, during transition time upper wavelet coefficients being affected: a significant coefficient is likely coming with additional significant coefficients at the same time location and lower scale level [3]. In this paper, two methods based on wavelet tree detection are presented, the obtained results are compared. We also present the results of sound classification method in noisy conditions.

2. SOUND EXTRACTION IN NOISY ENVIRONMENT

2.1 Noise and sounds

As no everyday life sound database was available in the scientific area, we have recorded a sound corpus. This corpus contains recordings made in the CLIPS laboratory, files of "Sound Scene Database in Real Acoustical Environment" (RCWP Japan) and files from a commercial CD: door slap, chair, step, electric shaver, hairdryer, door lock, dishes, glass breaking, object fall, screams, water, ringing, etc. The corpus contains 20 types of sounds with 10 to 300 repetitions per type. The test signal database has a duration of 3 hours and consists of 2376 files.

The sound classes of our corpus are described in the following table; the number of frames for each class is given too. Each frame has a duration of 16ms (256 samples at 16 kHz). Signal duration varies in a 500:1 ratio. Fast variations of the signal are related to short duration parts of the signal (some milliseconds).

Sound Class	Number of Frames (Entire corpus)	Duration (Each Sound)
Door Slap	47 398	375 ms
Breaking Glasses	9 338	15 ms-7.5 s
Ringing Phone	59 188	35 ms-10 s
Step Sound	3 648	1.4-5 s
Scream	17 509	0.37-5.8 s
Dishes Sounds	7943	125 ms-1.35 s
Door Lock	605	24 ms-117 ms

Table 1: Sound classes

This work is a part of the DESDHIS-ACI "Technologies for Health" project of the French Research Ministry. This project is a collaboration between the Clips ("*Communication Langagière et Interaction Personne-Système*") laboratory, in charge of the sound analysis, and the TIMC ("*Techniques de l'Imagerie, de la Modélisation et de la Cognition*") laboratory, charged with the medical sensors analysis and data fusion.

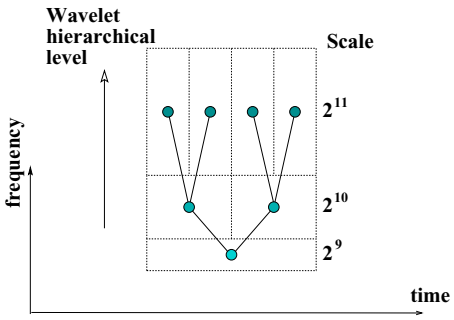

Figure 1: Tree of wavelet coefficients for N=2048 sample window (tree depth of 3 levels)

Two types of noise have been considered, the noise registered inside an experimental apartment[1], which is named HIS noise, and stationary white noise. HIS noise is a result of all noises in the building, he is a transient noise similar to usual sounds to detect, but transients are partially reduced by propagation inside the structure of the building. This kind of noise is not a stationary noise. First investigations showed that, unlike Dufaux studies [4], white noise performances are not sufficient to insure satisfactory performances in our actual case.

For this reason white noise study will only be used for literature result comparison, like in [4]. Evaluation of the algorithms has been made at 4 signal to noise ratios: 0, +10, +20 and +40dB.

2.2 Transients modeling

Methods based on wavelet transforms are often used for singularity characterization and transient detection, because of the compact support of wavelets in conjunction of the dyadic properties of these transforms. These two properties are allowing the analysis of reduced parts of the processing window. The figure 1 shows a wavelet tree with 3 level depth beginning at the highest hierarchical level. Each node is corresponding to a wavelet whose support is drawn in frequency and time domain. For wavelets of highest level the support in time is twice the sampling period.

For our purpose it is not necessary to determine the full tree corresponding to the transient, we limit our study to these 3 levels and we characterize each tree by his energy e, the sum of the energy of all nodes. We have chosen Daubechies wavelets ψ with 6 vanishing moments to compute DWT on 2048 sample windows (128 ms), the wavelet base is generated by translation and dilatation of the mother wavelet ψ [8]:

$$\left\{ \psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j} \right) \right\}_{(j,n)\in\mathbb{R}} \quad (1)$$

As we consider the energy e of the tree, the non significant nodes are implicitly not taken into account because they are negligible in the summation. With this approach the tree is not pruned and we don't eliminate nodes at scale 2^{11} if their mother node at scale 2^{10} is not significant, but this might not be very harmful because of the low depth of the tree.

A signal of chair falling with HIS noise is drawn on the bottom sub-figure of figure 2, the sound appears at time

[1]The HIS apartment is located in the TIMC laboratory building

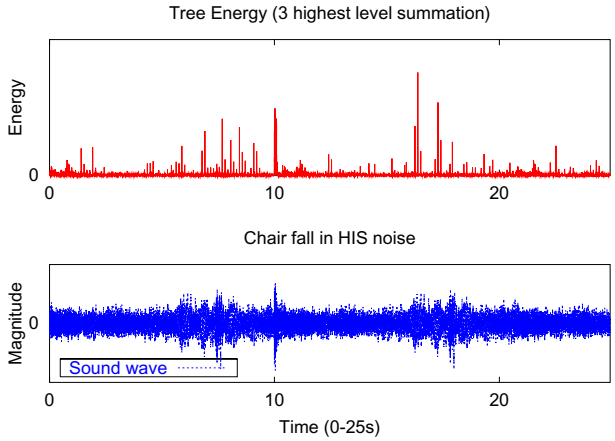

Figure 2: Sound signal and Tree Energy

$t = 10s$. The top sub-figure displays tree energy evolution across the time. Energy corresponding to useful signal is surrounded by isolated noise pulses which are sometimes greater but useful signal is associated with numerous adjacent trees and in this way could be detected.

2.3 Proposed detection algorithms

2.3.1 Several tree mean

DWT is calculated on $N = 2048$ sample windows (128ms) as shown in figure 3. From this DWT the energy e of each tree is obtained by time translation ($500\mu s$) across the transform. The means e_{means} of the 64 last values is calculated at each translation step in order to suppress noise influence. Since 16 kHz sampling rate, corresponding frame width is 32 ms. A transient is characterized by a large increase of e_{means}.

The detection th threshold is adaptive: $th = \kappa + 1.2 \cdot \mu_{e_{means}}$, with $\mu_{e_{means}}$ referring to the mean of the last values of e_{means} and κ to an adjusting parameter. The coefficient 1.2 was introduced because of remaining oscillations on e_{means}.

2.3.2 Threshold on the standard deviation

This algorithm (see flowchart in figure 3) is computing the DWT of consecutive $N = 2048$ sample windows. From this DWT the energy e of each tree is obtained as above by time translation across the transform. A median filter is applied to eliminate isolated trees which are only relevant of noise, the

Detection Method	SNR	HIS noise	White noise
Several tree mean	0dB	6.7%	5.9%
	$\geqslant$+10dB	0%	0%
Standard deviation	0dB	3%	22.7%
	$\geqslant$+10dB	0%	0%
Filtered energy	*0dB*	*71.3%*	*19.2%*
(conditioning	*+10dB*	*45.2%*	*6.1%*
median filter)	*+20dB*	*7.5%*	*6.1%*
	+40dB	*6.1%*	*6.1%*

Table 2: Detection EER, 198 tests at each SNR level (99 noised sounds, 99 pure noise)

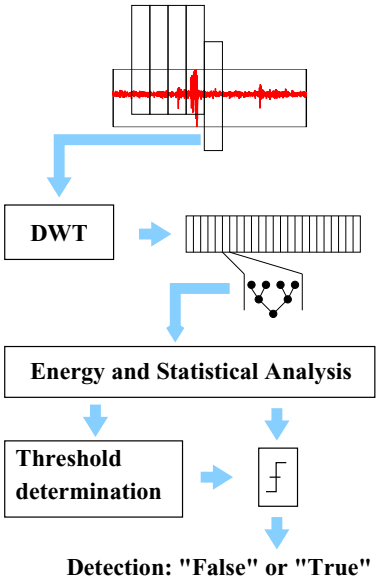

Figure 3: Detection algorithms using energy tree evaluation

Method	0dB	+10dB	+20dB	+40dB
Sev. tree mean	23.6ms	13.9ms	9ms	5.5ms
Standard dev.	30.6ms	13.4ms	11.5ms	8.4ms

Table 3: Mean of detection delay for sound duration shorter than 2s for HIS noise (78 tests at each SNR level)

width of the filter is 3. The standard deviation σ of the last 640 filtered energy values is calculated at each translation step: a high increase of the standard deviation is significant of a transient.

The detection is achieved by increase beyond an adaptive threshold $th = \kappa + \mu_{sigma}$, with μ_{sigma} referring to the mean of the last values of σ and κ to an adjusting parameter.

2.4 Detection results

Evaluation of each algorithm was done from COR curves giving *missed detection rate* (MDR) as function of *false detection rate* (FDR), the Equal Error Rate (EER) being achieved when MDR=FDR. Results for the two algorithms and for the conditioning median filtered energy described in [4] are given in table 2. Best results for HIS noise at 0dB SNR are obtained for *"Standard deviation"* (3%) and *"Several tree mean"* (6.7%), in the case of white noise *"Several tree mean"* (5.9%) is the best.

In order to insure best classification results, a short detection delay is very important. Delay means for the 2 proposed methods are given in table 3 at each SNR in the previous conditions (threshold choice in order to obtain Equal Error conditions) for sounds of short duration for which it is important to extract the most useful part of the signal. Best values at 0dB SNR are obtained for *"Several tree mean"*: 23.6ms; if SNR$\geqslant$+10dB they are below 14ms for the 2 methods. An additional part of signal may be added without critical incidence by deciding that signal is beginning 20 ms before detection time: it is needed neither to cut signal nor to transmit additional noise frames to the classification stage.

3. SOUND CLASSIFICATION

We have used a **G**aussian **M**ixture **M**odel (GMM) method in order to classify the sounds [9]. There are other possibility for the classification: HMM, Bayesian method, etc. GMM has been chosen because with other methods similar results has been obtained, although they are more complex.

3.1 Acoustical parameters

The first step of sound classification is acoustical parameters extraction. Acoustical parameters are a synthetic representation of time signal. Acoustical parameters classically used in speech/speaker recognition are: MFCC(Mel Frequencies Cepstral Coefficients), LFCC (Linear Frequencies Cepstral Coefficients), LPC(Linear Predictive Coefficients). Acoustical parameters used in speech/music/noise segmentation are : ZCR (zero crossing rate), RF (roll-off point), centroïd. **Zero Crossing Rate (ZCR)** is the number of crossings on time-domain through zero-voltage within an analysis frame. **Roll-off Point (RF)** is the frequency which is above 95% of the power spectrum. **Centroïd** represents the balancing point of the spectral power distribution within a frame.

3.2 GMM

The classification with a GMM method suppose that the acoustical parameters repartition for a sound class may be modeled with a sum of Gaussians. This method evolves in two steps: a training step and a classification step. In the training step for each sound class the Gaussian model is estimated. The training step start with a K-Means algorithms followed by EM algorithm(Expectation-Maximization) in 20 steps. In the classification step for each acoustical vector is calculated a likelihood for each sound class. The global likelihood for each class is the geometrical average of all acoustical vector likelihood. The signal belongs to the sound class for which likelihood is maximum.

3.2.1 Model Selection

The BIC (Bayesian Information Criterion) criterion is used in this paper in order to determinate the optimal number of Gaussians [10]. BIC criterion select the model trough the maximization of integrated likelihood: $BIC_{m,K} = -2.L_{m,K} + \nu_{m,K} \ln(n)$. Where $L_{m,K}$ is logarithmic maximum of likelihood, equal to $\log f(x|m,K,\widehat{\theta})$ (f is integrated likelihood), m is the model and K the component number of model, $\nu_{m,K}$ is the number of free parameters of model m and n is the number of frames. The minimum value of BIC indicate the best model.

The BIC criterion has been calculated for the sound class with the smallest number of files, for 2, 4, 5 and 8 Gaussians. The results of the table 4 are obtained for 16 MFCC parameters. Looking at the results, a number of Gaussians between 3 and 5 seem to correspond to the best sound modeling. We have decided to use 4 Gaussians.

No. of Gaussians	2	**3**	**4**	**5**	8
BIC	11043	**10752**	**10743**	**10757**	13373

Table 4: BIC for 2, 3, 4, 5 et 8 Gaussians (1577 tests)

	SNR [dB]				
Filtering	0	10	20	40	$\geqslant 55$
Without	48.3	27.2	13.1	11.1	10.1
With F1	40	20.5	14.6	10.4	10
With F2	40.4	20.9	15.1	10.7	10

Table 5: ECR for 16MFCC+ZCR+RF+Centroïd in the HIS noise presence (1577 tests for each SNR)

3.3 Noise attenuation

In order to increase the classification efficiency, wavelet filtering is applied before sound classification. The Wavelet Transform is more adapted to analyze and process impulsive signals than Fourier Transform which is adapted to periodical signals.

Two methods are tested on our test set. The general steps of the method are : DWT calculation on 256 samples window (9 wavelet coefficients), the application of thresholds on the DWT Coefficients, DWT inverse calculation.

Thresholds are applied to the absolute value of each Wavelet Transform coefficients. For the first method (F1) values under the threshold are cleared and other values are unmodified. For the second method (F2) values under the threshold are cleared; for other values a subtraction of estimated noise value is made ($B^i_{max}/10$). Threshold values for each DWT Coefficient are:

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

where T_i is the threshold applied to coefficient i of DWT and B^i_{max} the maximal value of coefficient i of DWT for the noise. The value B^i_{max} is estimated on the first 100ms of signal which are considered to contain only environmental noise.

This filtering threshold choice results from a study of the HIS noise and sounds. The sounds contain less useful information in the first five DWT coefficients, whereas in the case of HIS noise almost all information is located in low hierarchical level coefficients of DWT.

3.4 Classification results in noisy conditions

The sound classification is validated on the test set with 7 classes (the pure sounds and the sounds mixed with HIS noise at 0 dB, 10 dB, 20 dB and 40 dB of SNR). The sound classification performances are evaluated through the error classification rate (ECR) which represent the ratio between the bad classified sounds and the total number of sounds to be classified.

In the table 5 the classification results for 16 MFCC acoustical parameters coupled with zero crossing rate, Roll-off point and centroïd are presented. We can observe that for "pure" sounds we have 10% of classification error. In the noise conditions, the wavelet filtering give a gain, in absolute, of 8% for the ECR. The two methods of wavelet filtering has approximately same results.

4. CONCLUSION

We have presented detection and classification methods allowing us to detect and classify a sound event recorded in nursed home. Proposed detection method are resulting in low delay after signal beginning -typically 14 ms- so that link to classification step is not disturbed.

Detection is error-less for 10dB SNR and upper and error classification rate of 20% or better are reached in the same noise conditions; according to these two results we can conclude that this detection/classification system may be used under realistic conditions with moderate noise.

We are working to apply proposed detection techniques to speech recognition in order to allow call for help by the patient in our medical application.

These identification methods may have possible applications in multimedia classification or security sound surveillance.

REFERENCES

[1] G. Virone, D. Istrate, M. Vacher and all, "First Steps in Data Fusion between a Multichannel Audio Acquisition and an Information System for Home Healthcare," in *Proc. IEEE Engineering in Medicine and Biology Society*, Cancun, Mexico, Sept. 2003, pp. 1364–1367.

[2] L. Daudet, *Représentations structurelles de signaux audiophoniques - Méthodes hybrides pour des applications à la compression*. PhD Thesis, Marseille, 2000.

[3] L. Daudet, S. Mollat, and D. B. Torrésani, "Transient detection and encoding using wavelet coefficient trees," in *Proc. GRETSI 2001*, Toulouse, France, F. Flandrin Ed., Sept. 2001.

[4] A. Dufaux, L. Besacier, M. Ansorge and F. Pellantini, "Automatic Sound Detection and Recognition for Noisy Environment," in *EUSIPCO 2000*, Tampere, Finland, Sept. 2000.

[5] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Journal of Signal Processing, Special issue on Image and Video Coding Beyond Standards*, vol. 82(11), pp. 1595-1617, Nov. 2002.

[6] M.Cowling, and R. Sitte, "Analysis of Speech Recognition Techniques for use in a Non-Speech Sound Recognition System," in *Proc. Digital Signal Processing for Communication Systems*, Jan. 2002.

[7] L. Lu, H.J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transaction on Speech and Audio Processing*, vol. 10(7), pp. 504-516, Jan. 2002.

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

[9] D. Reynolds, *Speaker Identification and Verification using Gaussian Mixture Speaker Models*, Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland,pp 27-30, 1994.

[10] G. Schwarz, *Estimating the dimension of a model*, Annals of Statistics,1978,pp.461-464.

A.2 Article IEEE Trans. on Information technology in Biomedicine 2006

"Information Extraction From Sound for Medical Telemonitoring"

D. ISTRATE, E. CASTELLI, M. VACHER, L. BESACIER, J.-F. SERIGNAT

IEEE Transactions on Information Technology in Biomedicine, April 2006, vol.**10**, issue 2, pp. 264-274, ISSN : 1089-7771.

A.3 Article IEEE Trans. on Information technology in Biomedicine 2010

"SVM-Based Multi-Modal Classification of Activities of Daily Living in Health Smart Homes : Sensors, Algorithms and First Experimental Results"

A. FLEURY, M. VACHER, N. NOURY

IEEE Transactions on Information Technology in Biomedicine, March 2010, vol. **14**, nb. 2, pp. 274 -283.

A.4 Article E-Health and Medical Communications 2011

"Development of Audio Sensing Technology for Ambient Assisted Living : Applications and Challenges"

M. VACHER, F. PORTET, A. FLEURY, N. NOURY

International Journal of E-Health and Medical Communications, vol. **2**(1), pp. 35-54, January-March 2011.

A.5 Conférence Interspeech 2011

"Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions"

B. LECOUTEUX, M. VACHER, F. PORTET

Interspeech 2011, Florence, Italy, Aug. 2011, 4p.

Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions

Benjamin Lecouteux, Michel Vacher, François Portet

Laboratoire d'Informatique de Grenoble, GETALP Team
UMR CNRS/UJF/G-INP 5217, Grenoble, France
Benjamin.Lecouteux@imag.fr, Michel.Vacher@imag.fr, Francois.Portet@imag.fr

Abstract

While the smart home domain has become a major field of application of ICT to improve support and wellness of people in loss of autonomy, speech technology in smart home has, comparatively to other ICTs, received limited attention. This paper presents the SWEET-HOME project whose aim is to make it possible for frail persons to control their domestic environment through voice interfaces. Several state-of-the-art and novel ASR techniques were evaluated on realistic data acquired in a multiroom smart home. This *distant speech* French corpus was recorded with 21 speakers playing scenarios including activities of daily living in a smart home equipped with several microphones. Techniques acting at the decoding stage and using *a priori* knowledge such as DDA give better results (WER=8.8%, Domotic F-measure=96.8%) than the baseline (WER=18.3%, Domotic F-measure=89.2%) and other approaches.

Index Terms: home automation, smart home, distant speech, multisource ASRs, keyword detection

1. Introduction

Since the rise of *Ubiquitous Computing*, new ways of conceiving our home environment appeared. One of these, is the development of *smart homes* which are habitations equipped with a set of sensors, actuators, automated devices and centralised software which control the increasing amount of household appliances ranging from lights, automatic motorised blinds... to Hi-Fi systems, PCs, alarms systems, etc. that fit modern homes. These smart homes represent a promising solution to support the elderly and disabled persons in living in their own home as autonomously as possible. Among all the interaction and sensing technologies used in smart home, speech processing technology has a great potential to become one of the major interaction modalities in smart home. Indeed, voice interfaces are much more adapted to disabled people and the ageing population who have difficulties in moving or seeing, than tactile interfaces (e.g., remote control) which require physical and visual interaction [1, 2]. Moreover, voice command is particularly suited to distress situations. A person, who cannot move after a fall but being conscious, may have still the possibility to call for assistance while a remote control may be unreachable. Despite all this, very few smart home projects have seriously considered speech recognition in their design [1, 2]. Part of this can be attributed to the complexity of setting up this technology in a real environment and to important challenges that still need to be overcome [3].

The SWEET-HOME project has started in 2010 to address some of these challenges. One of the major issues that prevents the development of speech technology in real home setting is the poor performance of Automatic Speech Recognition (ASR) in noisy environment [3]. Indeed, ASR systems have reached correct performances with close talking microphones (e.g. head-set), but the performance decreases significantly as soon as the microphone is moved away from the mouth of the speaker. In realistic conditions, this deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise such as TV, radio and devices [4]. All these problems, related to the so called *'distant speech'* context, should be taken into account in the home context [5]. While, user linguistic preferences, dialogues and age dependant voice interfaces have been studied during this decade [1, 2, 6], distant speech in smart home received attention very recently within the speech processing community [7].

This paper presents results of state-of-the-art and novel ASR techniques evaluated on realistic data acquired in a multiroom smart home. Before presenting our experimental framework (Section 3), the proposed techniques are described in Section 2. The experiments and results are then presented in Section 4. This paper concludes with brief remarks about the results and future work.

2. SWEET-HOME project and corpus

The SWEET-HOME project (`sweet-home.imag.fr`) aims at designing a new smart home system based on audio technology focusing on three main aspects: to provide assistance via *natural man-machine interaction* (voice and tactile command), to ease *social e-inclusion* and to provide *security reassurance* by detecting situations of distress. If these aims are achieved, then the person will be able to pilot, from anywhere in the house, their environment at any time in the most natural way possible. The targeted smart environments in which speech recognition must be performed thus include multi-room homes with one or more microphones per room set near the ceiling. This places the project in a distant-speech context where microphones may be far apart from each other and may thus record similar or very different sources. The most close projects seem to have focused mainly on one-room microphone array [1] or one or unspecified number of microphones [2, 6].

To achieve the project goals, the DOMUS smart home depicted in Figure 1 was adapted to acquire a realistic corpus and to test the developed techniques. This smart home was set up by the Multicom team of the Laboratory of Informatics of Grenoble, partner of the project. It is a thirty square meters suite flat including a bathroom, a kitchen, a bedroom and a study, all equipped with sensors and effectors such as infra-red presence detectors, contact sensors, video cameras (used only for annotation purpose), etc. In addition, seven microphones were set in the ceiling.

An experiment was conducted to acquire a representative speech corpus composed of utterances of domotic order, distress call and casual sentence. This corpus is called the SWEET-

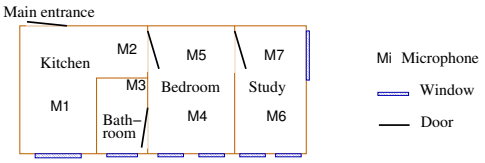

Figure 1: Position of the 7 microphones in the DOMUS Home

HOME *speech corpus*. 21 persons (including 7 women) participated to a 2-phase experiment to record, among other data, speech corpus in a daily living context. The average age of the participants was 38.5 ± 13 years (22-63, min-max). To ensure that the audio data acquired would be as close as possible to real daily living sounds, the participants were asked to perform several daily living activities in the smart home. A visit, before the experiment, was organized to make sure that the participants will find all the needed items to perform the activities. No instruction was given to any participant about either how they should speak or in which direction. Consequently, no participant emitted sentences directing their voice to a particular microphone. The distance between the speaker and the closest microphone is about 2 meters. Sound data were recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card [3].

The first phase (**Phase 1**) consisted in following a scenario of activities without condition on the time spent and the manner of achieving them (having a breakfast, simulate a shower, get some sleep, clean up the flat using the vacuum, etc.). During this first phase, participants uttered 40 predefined casual sentences on the phone (e.g., "Allo" (*Hello*), "J'ai eu du mal à dormir" (*I slept badly*)) but were also free to utter any sentence they wanted (some did speak to themselves aloud). The second phase (**Phase 2**) consisted in reading aloud a list of 44 sentences whose 9 were distress sentences (e.g., "À l'aide"(*help*), "Appelez un docteur" (*call a doctor*)) and 3 were domotic orders (e.g., "Allumez la lumière" (*turn on the light*)). In this paper, experiments are performed without device noise (TV, radio, vacuum, ..).

Finally, the French SWEET-HOME speech corpus is made of 862 sentences (38 minutes 46s per channel in total) for **Phase 1**, and 917 sentences (40 minutes 27s per channel in total) for **Phase 2** all from 21 speakers. Each sentence is available for each channel and has been humanly annotated on the best Signal-to-Noise Ratio (SNR) channel. The average SNR for the considered sentences of the best SNR channel is 20.3 dB (SNR is typically around 55 db in studio record). It must be clear that the data from the 7 microphones was the only data source used in this study.

3. Proposed approaches for robust ASR

To detect domotic commands in the SWEET-HOME context, we propose a three-stage approach. The first one detects audio activity and classifies it as speech or other sound, the second one extracts the utterances using an ASR system and the last one recognizes a vocal command or a distress situation from the decoded utterances. This paper describes the two last stages, for the first stage the reader is referred to [4].

To address the issues of the SWEET-HOME context (noise, distant-speech) and to benefit from it (multiple microphones which are continuously recording) we propose to test the impact of some state-of-the-art and novel techniques that fuse the streams of information at three independent levels of the speech processing: *acoustic signal enhancement*, *decoding enhancement*, and *ASRs output combination*. The remaining of this section presents the implemented techniques for robust ASR and the chosen method for vocal order recognition.

3.1. Beamforming

At the acoustic level, it may be interesting to fuse the different channels in order to enhance the signal. However, a simple sum of signals would result in a worse single channel with echoes. That is why a beam-forming algorithm [8] was used to merge all channels in a single one which fed an ASR system. Beamforming involves low computational cost and combines efficiently acoustic streams to build an enhanced acoustic signal.

The acoustic beamforming algorithm is based on the *weighted&sum microphone array* theory. Given M microphones, the signal output $y[t]$ is computed by:

$$y[t] = \sum_{m=1}^{M} W_m[t] x_m[t - D^{(m,ref)}[t]]$$

where $W_m[t]$ is the weight for microphone m at time t, knowing that $\sum_{m=1}^{M} W_m[t] = 1$, the signal of the m^{th} channel is $x_m[t]$ and $D^{(m,ref)}[t]$ is the delay between the m^{th} channel and the reference channel. In our experiments, the reference channel was the one with the highest SNR overall in **Phase 2** and the 7 signals were entirely combined for each speaker rather than doing a sentences based combination (the tested algorithm failed with too short sentences). Once the new signal y is computed, it can feed a monosource ASR stage.

3.2. Driven Decoding Algorithm

At the decoding level, a novel version of the Driven Decoding Algorithm (DDA) was applied. DDA aims to align and correct auxiliary transcripts by using a speech recognition engine [9, 10]. This algorithm improves system performance dramatically by taking advantage of the availability of the auxiliary transcripts.

DDA acts at each new generated assumption of the ASR system. The current ASR assumption is aligned with the auxiliary transcript (from a previous decoding pass). Then a matching score α is computed and integrated with the language model [9]:

$$\tilde{P}(w_i|w_{i-1}, w_{i-2}) = P^{1-\alpha}(w_i|w_{i-1}, w_{i-2})$$

where $\tilde{P}(w_i|w_{i-1}, w_{i-2})$ is the updated trigram probability of the word w_i knowing the history w_{i-2}, w_{i-3}, and $P(w_i|w_{i-1}, w_{i-2})$ is the initial probability of the trigram.

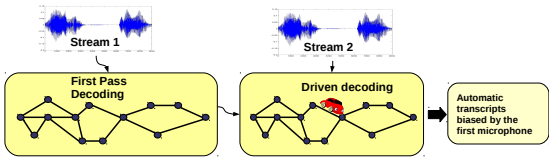

Figure 2: **DDA** used with two streams: the first stream allows one to drive the second stream

We propose to use a variant of the Driven Decoding Algorithm where the output of the first microphone is used to drive the output of the second one (cf. Figure 2). This approach presents two main benefits:

- The second ASR system speed is boosted by the approximated transcript (only 0.1xRT),

- DDA merges truly and easily the information from the two streams while voting strategies (such as ROVER) do not merge ASR systems outputs.

The applied strategy is dynamic and used, for each utterance to decode, the best channel for the first pass and the second best channel for the last pass.

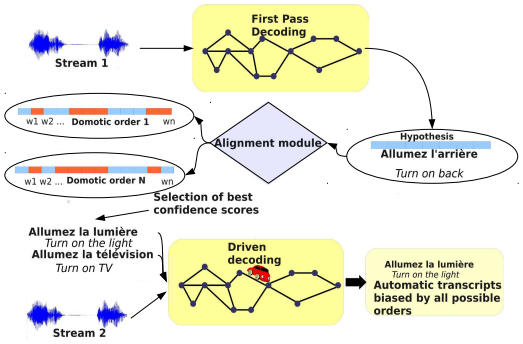

Figure 3: **DDA 2-level**: vocal orders are recognized from the first decoded stream which are then used to drive the decoding of the second stream

This approach was extended to take into account *a priori* knowledge about the expected utterances. The ASR system is driven by vocal orders recognized during the first pass. This method is called **DDA 2-level**: speech segments of the first pass are projected into the $3 - best$ vocal orders by using an edit distance (cf. 3.4) and injected via DDA into the ASR system for the fast second pass as presented in Figure 3.

3.3. ROVER

At the ASR combination level, a ROVER [11] was applied. ROVER is expected to improve the recognition results by providing the best agreement between the most reliable sources. It combines systems output into a single word transition network. Then, each branching point is evaluated with a vote scheme. The word with the best score is selected (number of votes weighted by confidence measures). This approach necessitates high computational resources when several sources need to be combined and real time is needed (in our case, 7 ASR systems must operate concurrently).

A baseline ROVER was tested using all available channels without *a priori* knowledge. In a second time, an *a priori* confidence measure based on the SNR was used: for each decoded segment s_i from the i^{th} ASR system, the associated confidence score $\phi(s_i)$ was computed by $\phi(s_i) = 2^{R(s_i)}/\sum_{j=1}^{7} 2^{R(s_j)}$ where $R()$ is the function computing the SNR of a segment and s_i is the segment generated by the i^{th} ASR system. For each annotated sentence a silence period I_{sil} at the beginning and the end is taken around the speech signal period I_{speech}. The SNR is thus evaluated as:

$$R(S) = 10 * log(\frac{\sum_{n \in I_{speech}} S[n]^2}{|I_{speech}|} / \frac{\sum_{n \in I_{sil}} S[n]^2}{|I_{sil}|}).$$

Finally, a ROVER using only the two best channels overall was tested in order to check whether other channels contain redundant information and whether good results can be reach with reasonable computational cost.

3.4. Detection of domotic orders and distress sentences with proposed approaches

We propose to transcript each domotic order and distress sentences in a phoneme graph in which each path corresponds to a variant of pronunciation. Then the number of sentences to detect is 12 (3 domotic orders + 9 distress sentences). Automatic transcripts are transcribed in the same way.

In order to locate domotic orders into automatic transcripts T of size m, each sentence of size n from domotic orders H are aligned to T by using a Dynamic Time Warping (DTW [12]) algorithm at the phonetic level. The deletion, insertion and substitution costs were computed empirically. The cumulative distance $\gamma(i,j)$ between H_j and T_i is computed as: $\gamma(i,j) = d(T_i, H_j) + min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\}$

Each domotic order is aligned and associated with an alignment score: the percentage of well aligned symbols. The domotic order with the best score is then selected for decision according a detection threshold. This approach takes into account some recognition errors such as word endings or slight variations. Moreover, in many cases, a miss-decoded word is phonetically close from the good one (due to the close pronunciation).

4. Experiments and results

In all experiments, the **Phase 1** corpus was used for development and training whereas the **Phase 2** corpus served for the evaluation. This section presents the ASR tuning and the experimental results of the proposed approaches.

4.1. The Speeral ASR system

The LIA (Laboratoire d'Informatique d'Avignon) speech recognition tool-kit Speeral [13] was chosen as unique ASR system. This choice was made based on experiments we undertook with several state-of-the-art ASR systems and on the fact that DDA is only implemented in Speeral. Speeral relies on an A^* decoder with HMM-based context-dependent acoustic models and trigram language models. HMMs are classical three-state left-right models and state tying is achieved by using decision trees. Acoustic vectors are composed of 12 PLP (Perceptual Linear Predictive) coefficients, the energy, and the first and second order derivatives of these 13 parameters.

In the study, the acoustic models were trained on about 80 hours of annotated speech. Given the targeted application of SWEET-HOME the computation time should not be a breach of real-time use. Thus, the 1xRT Speeral configuration was used. In this case, the time required by the system to decode one hour of speech signal is real-time (noted 1xRT). The 1xRT system uses a strict pruning scheme. Furthermore, acoustic models were adapted for each of the 21 speaker by using the Maximum Likelihood Linear Regression (MLLR) and the annotated **Phase 1** corpus. MLLR adaptation is a good compromise while only a small amount of annotated data is available

For the decoding, a 3-gram language model (LM) with a 10K lexicon was used. It results from the interpolation of a generic LM (weight 10%) and a specialized LM (weight 90%). The *generic* LM was estimated on about 1000M of words from the French newspapers *Le Monde* and *Gigaword*. The *specialized* LM was estimated from the sentences (about 200 words) that the 21 participants had to utter during the experiment (domotic orders, casual phrases, etc.).

4.2. Results

Results of the approaches are presented table 1. The ASR stage was evaluated using the Word Error Rate (WER) whereas the vocal order recognition (classification) stage was evaluated using recall/precision/F-measure triplet: the number of domotic orders is about 10. Domotic orders were manually specified by annotating all sentences. During the detection, if a marked do-

motic order is well detected, it is considered as detected. In all other cases, a detected order is considered as false detection. For each approach, the presented results are the average over the 21 speakers (plus standard deviation for the WER). For the sake of comparison, results of a baseline and an oracle baseline systems are provided. The baseline system outputs the best decoding amongst 7 ASR systems according to the highest SNR. The oracle baseline is computed by selecting the best WER for each speaker.

Method	WER$^{\pm SD}$	Domotic recall	Domotic precision	F -measure
Baseline	$18.3^{\pm 12.1}$	88.0	90.5	89.2
Oracle Baseline	$17.7^{\pm 10.3}$	88.5	91.3	89.9
Beam Forming	$16.8^{\pm 8.3}$	89.0	92.6	90.8
DDA +SNR	$11.4^{\pm 5.6}$	93.3	97.3	95.3
DDA 2 lev.+SNR	$\mathbf{8.8}^{\pm 3.7}$	**95.6**	**98.1**	**96.8**
ROVER	$20.6^{\pm 8.5}$	85.0	90.0	87.4
ROVER 2c+SNR	$13.0^{\pm 6.6}$	91.3	95.3	93.3
ROVER +SNR	$\mathbf{12.2}^{\pm 6.1}$	**92.7**	**97.4**	**95.0**
ROVER Oracle	$7.8^{\pm 2.7}$	99.4	98.9	99.1

Table 1: WER, Domotic orders detection

The baseline system achieved a 18.3 % WER (best SNR channel). All proposed SNR-based approaches benefited from the multiple available microphones. Beamforming led to a 8.1% relative WER improvement. This result shows that combining all channels increases the ASR task robustness. The DDA method showed a 37.8% relative improvement by using the SNR. The 2 level DDA presented a 52 % relative improvement with a very high stability (SD=3.7): this gain is easily explained as the second decoding pass was perfused with *a priori* knowledge (i.e., domotic orders) triggered by the first pass. Finally, the SNR-based ROVER led to a 33.4% relative improvement.

In all configurations, accuracy of the vocal orders recognition was good: the baseline recognition gave a 89.2% F-measure. It can be observed that in other configurations the spotting task correlated with the WER. Thereby, ROVER and the two DDA configurations led to a significant F-measure improvement over the baseline of about 7% absolute. Beamforming gain was not significant. ROVER performed detections similar to the DDA approaches, but required to decode all channels. Finally, the best configuration was based on the 2 level DDA leading to a 96.8% F-measure.

5. Conclusion

Several approaches were presented to perform accurate vocal order recognition in multi-room smart-homes where audio information is captured by several microphones in a distant speech context. The proposed approaches were acting at the three main levels of the ASR task: acoustic, decoding and hypothesis selection. Some of them included *a priori* knowledge either dynamically computed such as the SNR or acquired off line such as the predefined domotic orders.

Results confirmed that the use of the seven microphones improved the ASR accuracy. Beamforming improved the WER (16.8%), however its performance were very close to the baseline one (18.3%). This may be due to the fact that the seven microphones are far apart from each other and might not contain enough redundancy to obtain a really enhanced acoustic signal. The Driven Decoding Algorithm gave the best performance with a 11.4% WER and 95.3% F-measure for vocal order classification. DDA results were only slightly better than the ROVER results, however DDA needs only two channels while ROVER necessitates 7 ASR systems performing concurrently to approach DDA performances. The DDA computational cost is thus very low compared to the ROVER one. Moreover, the 2-level DDA approach makes it possible to include *a priori* knowledge to increase performances to 8.8% WER and 96.8% F-Measure with much better stability than the baseline (3.7% WER standard deviation vs. 10.3%). However, this amelioration will be achieved only if test data contains domotic orders. This study shows that good recognition rate can be obtained by adapting classical ASR systems mixing multisource and domain knowledge. We plan to adapt these approaches to noisy conditions notably by applying source separation techniques to real daily living records composed of uncontrolled noise.

6. Acknowledgements

This work is a part of the SWEET-HOME project founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-09-VERS-011)

7. References

[1] A. Vovos, B. Kladis, and N. Fakotakis, "Speech operated smart-home control system for users with special needs," in *Proc. InterSpeech 2005*, 2005, pp. 193–196.

[2] M. Hamill, V. Young, J. Boger, and A. Mihailidis, "Development of an automated speech recognition interface for personal emergency response systems," *Journal of NeuroEngineering and Rehabilitation*, vol. 6, 2009.

[3] M. Vacher, F. Portet, A. Fleury, and N. Noury, "Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges," *International Journal of E-Health and Medical Communications*, vol. 2, no. 1, pp. 35–54, 2011.

[4] M. Vacher, A. Fleury, J.-F. Serignat, N. Noury, and H. Glasson, "Preliminary Evaluation of Speech/Sound Recognition for Telemedicine Application in a Real Environment," in *Proc. InterSpeech 2008*, 2008, pp. 496–499.

[5] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Published by Wiley, 2009.

[6] R. C. Vipperla, M. Wolters, K. Georgila, and S. Renals, "Speech input from older users in smart environments: Challenges and perspectives," in *HCI International: Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, 2009.

[7] J. Barker, H. Christensen, N. Ma, P. Green, and E. Vincent, "The PASCAL 'CHiME' Speech Separation and Recognition Challenge," in *InterSpeech 2011*, 2011, (to appear).

[8] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[9] B. Lecouteux, G. Linarès, J. Bonastre, and P. Nocéra, "Imperfect transcript driven speech recognition," in *Proc. InterSpeech'06*, 2006, pp. 1626–1629.

[10] B. Lecouteux, G. Linarès, Y. Estève, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Proc. IEEE ICASSP 2008*, 2008, pp. 1549–1552.

[11] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop ASRU*, 1997, pp. 347–354.

[12] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Workshop on Knowledge Discovery in Databases (KDD'94)*, 1994, pp. 359–370.

[13] G. Linarès, P. Nocéra, D. Massonié, and D. Matrouf, "The LIA speech recognition system: from 10xRT to 1xRT," in *Proc. TSD'07*, 2007, pp. 302–308.

ANNEXE B

Curriculum Vitae

Michel Vacher
56 ans
Ingénieur de Recherche CNRS
Laboratoire d'Informatique de Grenoble, équipe GETALP
UJF BP53 38041 Grenoble Cedex 9
Tel : 0476635795

Michel.Vacher@imag.fr

Curriculum Vitae

Ce document est une synthèse de mon parcours et comprend, outre un résumé succinct des positions professionnelles que j'ai occupées et de mes diplômes (section 1), une description détaillée des encadrements d'étudiants réalisés (section 2), des projets nationaux que j'ai encadrés ou auxquels j'ai participé (section 3). Il dresse ensuite une liste de mes activités scientifiques (section 4) et de mes publications (section 5).

De par ma formation, mes premières activités de recherche ont été tournées vers le traitement du signal et la physique. Mes travaux de thèse étaient orientés en ce sens et comprenaient une part d'expérimentation physique et de traitement du signal, ainsi qu'une part importante consacrée à la simulation du mouvement d'une bulle de gaz dans un liquide, ce qui m'a permis de calculer le spectre d'émission ultrasonore d'une bulle de gaz. J'ai ensuite travaillé sur la simulation et le traitement d'images en microscopie à haute résolution, essentiellement dans le cadre de l'étude des quasi-cristaux à symétrie décagonale et du désordre des réseaux de Vortex bidimensionnels sur des surfaces de supraconducteurs à haute température critique.

J'ai ensuite effectué une reconversion thématique vers l'informatique lors de mon arrivée dans le Groupe d'Étude de l'Oral et du Dialogue (GEOD) du laboratoire CLIPS. Je me suis alors orienté vers l'analyse sonore dans les habitats intelligents. Cette problématique nécessite de progresser sur les théories de reconnaissance des signaux complexes dans un environnement sonore perturbé. Il s'agit en particulier de mettre en œuvre et d'adapter des techniques sophistiquées de traitement du signal (détection, déréverbération, séparation de sources, ...) afin de permettre ensuite la recherche d'informations au niveau sonore, c'est à dire l'identification des sons de la vie courante et la reconnaissance des paroles prononcées. Ces informations sont ensuite réunies avec des informations issues d'autres modalités pour détecter, identifier, interpréter des situations à un plus haut niveau (localisation ou activité de la personne), ce qui est nécessaire pour finalement prendre une décision informée et optimale. Mes travaux concernent plusieurs branches de l'informatique et présentent un aspect multidisciplinaire.

Table des matières

1 Curriculum Vitae résumé

1.1 Positions occupées

- **Ingénieur de Recherche CNRS au Laboratoire d'informatique de Grenoble** au sein du Groupe d'Étude pour la Traduction Automatique et le Traitement Automatisé des Langues et de la Parole depuis 2007.
- **Ingénieur de Recherche CNRS au laboratoire CLIPS** (Communication Langagière et Interaction Personne Système) au sein de l'équipe GEOD (Groupe d'Etude sur l'Oral et le Dialogue), 2001-2006.
- **Ingénieur de Recherche CNRS au LTPCM** (Laboratoire de Thermodymamique et Physico-Chimie Métallurgique) au sein de l'équipe Physique du Métal, 1986-2000.
- **Ingénieur, chef de projet** dans l'industrie, 1982-1986.
- **Attaché de recherche** sous contrat au Laboratoire de Traitement du Signal et d'Ultrasons de l'INSA de Lyon, 1979-1981.

1.2 Diplômes

- **Doctorat d'ingénieur** (1982), INSA de Lyon
  Titre de la thèse : *Étude du comportement non linéaire de bulles de gaz au voisinage de leur résonance dans un champ ultrasonore : application à la détection, à la localisation et au calibrage de bulles stationnaires dans un liquide*, sous la direction du professeur R. Goutte, 189 p.
- **DEA** (1980), spécialité Acoustique, INSA de Lyon, *Détection de bulles de gaz stationnaires par mesure de l'harmonique 2 d'une onde ultrasonore*.
- **Institut National des Sciences Appliquées de Lyon** (1979), département Génie Électrique option Télécommunications.

1.3 Domaines de recherche (mots-clefs)

- Traitement des signaux audio et de parole
  - Extraction et analyse d'éléments non linguistiques (sons)
  - Reconnaissance automatique de la parole et de la parole distante
  - Analyse du signal de parole dans le bruit

- Analyse de données multimodales
  - Reconnaissance d'activités de la vie quotidienne
  - Localisation de la personne

2 Activités d'encadrement

2.1 Résumé

	Thèses	**Mémoires CNAM**	**DEA ou Master-R**	**PFE**
Encadrement total		2 soutenus		2 soutenus
Co-encadrement	1 soutenue (EEATS) 2 en cours (MSTII)	1 soutenu	2 soutenus (EEATS) 1 soutenu (IDL - Stendhal)	4 soutenus

2.2 Thèses de doctorats

- A. Fleury : *Détection de motifs temporels dans les environnements multiperceptifs. Application à la classification automatique des Activités de la Vie Quotidienne d'une personne suivie à domicile en télémédecine.* Doctorat de l'Université Joseph Fourier, école doctorale EEATS (Électronique, Électrotechnique, Automatique & Traitement du Signal), **thèse soutenue** en octobre 2008.
- P. Chahuara : *Contrôle intelligent de la domotique à partir d'informations temporelles multi sources imprécises et incertaines*, Université de Grenoble, école doctorale MSTII (Mathématiques, Sciences et Technologies de l'Information et de l'Informatique), soutenance prévue en 2012.
- F. Aman : *Reconnaissance automatique de la parole des personnes âgées pour les services d'assistance aux personnes à domicile*, Université de Grenoble, école doctorale MSTII (Mathématiques, Sciences et Technologies de l'Information et de l'Informatique), soutenance prévue en 2013.

2.3 Mémoires CNAM

- S. Chaillol : *Détection et classification de sons dans un Habitat Intelligent pour la Santé*, mémoire CNAM, soutenu en décembre 2006.
- H. Glasson : *Système autonome d'acquisition temps réel multivoie de signaux sonores dans un Habitat Intelligent Santé (HIS)*, mémoire CNAM, soutenu en juin 2008.
- S. Méniard : *Système autonome d'acquisition temps réel multivoie de signaux sonores et de parole dans un Habitat Intelligent*, mémoire CNAM, soutenu le 10 juin 2011.

2.4 Master Recherche

- N. Gac : *Classification Parole/Sons environnementaux*, Master 2 recherche SIPT/EEATS, juillet 2004.
- N. Guirand : *Suppression de sources sonores parasites dans un environnement Habitat Intelligent pour la Santé (HIS)*, Master 2 recherche SIPT/EEATS, juin 2008.
- A. Maatallaoui : *Classification des sons de la vie courante dans un bâtiment intelligent en utilisant des méthodes probabilistes ou hiérarchiques*, Master 2 recherche IDL (Stendhal), soutenu le 24 juin 2011.

2.5 Projets de Fin d'Étude

- R. Dugheanu : *Évaluation des outils de reconnaissance de parole dans le cas de voix de personnes âgées*, Master professionnel IDL, soutenu le 24 juin 2011.
- C. Debeaux : *Implémentation d'un système de reconnaissance de la parole sur un système embarqué*, Licence professionnelle, IUT de Bourg en Bresse-Lyon 1, soutenance en septembre 2010.

- D. Cristobal Canals : *Optimisation de l'acquisition de sons et de parole sur une plate-forme de télé-surveillance médicale*, Enserg, soutenance en juin 2008.
- M. Bobu : *Optimisation de la classification des sons sur une plate-forme de télésurveillance médicale*, Université "Politehnica" de Bucarest, juin 2007.
- J. Torras : *Algorithme de suppression de sources sonores parasite dans un environnement Habitat Intelligent pour la Santé*, Enserg, soutenance en juin 2007.
- P. Menendez Garcia : *Plate-forme de démonstration pour la surveillance sonore d'un espace perceptif*, Ensimag, soutenance en juin 2005.

3 Responsabilité de projets nationaux

3.1 Projets achevés

RESIDE-HIS (IMAG 2000-2001) et **DESDHIS** (ACI Santé du Ministère de la Recherche 2002-2004) : J'ai participé de manière active à ces 2 projets lors de mon arrivée dans l'équipe GEOD. Le but et l'originalité de ces 2 projets consistaient à aider à la détermination d'une situation éventuelle de détresse d'un patient à son domicile à partir de l'analyse des sons et de la parole émis dans chacune des pièces de l'appartement. En effet, la reconnaissance de certains sons critiques ainsi que de mots-clefs caractéristiques d'un appel à l'aide, de cris ou de gémissements peut apporter une information très intéressante et primordiale et éviter d'utiliser une caméra vidéo qui pourrait être mal perçue par le patient. Je me suis particulièrement intéressé à la détection de signal sonore dans le bruit, à la classification des sons de la vie courante et à la discrimination entre sons de la vie courante et parole.

DESDHIS2 (2005-2008) collaboration avec le laboratoire TIMC-IMAG (coresponsabilité avec N. Noury) au travers d'une thèse (bourse ministère) que j'ai coencadrée sur l'analyse multimodale en vue de l'assistance à domicile. Le dispositif étudié prend place au sein d'un habitat dit "intelligent" et a pour but, à terme, de contribuer au maintien de patients à leur domicile en transmettant les alarmes vers un proche ou des centres de médicosurveillance en passant par le filtre d'un système opérant la fusion de données avec des capteurs d'activité et des capteurs médicaux.

Je me suis particulièrement intéressé à la reconnaissance des appels de détresse de la personne dans l'appartement et à la reconnaissance des activités de la vie quotidienne (AVQ ou ADL) en prenant en compte les données issues des microphones et de capteurs placés dans l'appartement de test ainsi que d'un capteur porté par la personne.

CIRDO-Formation (2009-2010) Réponse à l'appel "Services à la personne : innover pour développer l'offre de services" du Ministère de l'Emploi - projet CIRDO-Formation. Le projet était porté par le CATEL, j'étais le responsable local pour GETALP. Le montant de l'aide était pour GETALP de 10 000€.

Le but du projet est de permettre une meilleure efficacité du service d'aide à la personne et ainsi de lutter contre la dénutrition des personnes âgées. L'objectif était double : (1) étude bibliographique sur la reconnaissance de la parole des personnes âgées, et, (2) faisabilité de l'implémentation d'un système de reconnaissance sur le dispositif.

3.2 Projets en cours

Sweet-Home ANR VERSO 2009 (Réseaux du Futur et Services), durée 3 ans.
Je suis responsable du projet Sweet-Home auquel collaborent les équipes GETALP et MULTICOM du LIG, l'équipe ANASON de l'ESIGETEL et les sociétés Theoris, Technosens et Camera-Contact. Ce projet finance, en ce qui concerne les partenaires académiques, un doctorant et un post-doctorant pour GETALP, ainsi qu'un doctorant pour ESIGETEL. Personnel permanent LIG : 74 hommes.mois, personnel temporaire LIG : 48 hommes.mois, aide allouée au LIG : 277 486€. Labellisation par les pôles de compétitivité Minalogic et Cap-Digital.

L'évolution des technologies de la communication a favorisé l'émergence de nouvelles façons de concevoir un habitat, d'où le concept d'Habitat Intelligent. La Maison Intelligente est une résidence équipée de technologie informatique qui anticipe et répond aux besoins de ses occupants en essayant de gérer de manière optimale leur confort et leur sécurité par action sur la maison, et en mettant en œuvre des connexions avec le monde extérieur. Cependant, ces Maisons Intelligentes tendent à être équipées de dispositifs dotés d'interfaces de plus en plus complexes et d'autant plus difficiles à maîtriser par l'usager. Les personnes qui bénéficieraient le plus de ces nouvelles technologies sont les personnes en perte d'autonomie, les personnes atteintes de handicaps moteur ou fragilisées par des pathologies diverses. Elles sont, de plus, les moins aptes à utiliser des interfaces complexes étant donné leur handicap ou leur manque de familiarité avec les nouvelles technologies. Il devient donc indispensable de prévoir une assistance facilitant la vie quotidienne et l'accès à l'ensemble des systèmes dits "domotiques" au travers de l'Habitation Intelligente. Les interfaces tactiles usuelles devront être complétées par des interfaces plus accessibles, ne sollicitant ni la vue, ni le mouvement, grâce notamment à un système réactif à la parole ; elles trouveront également leur utilité lorsque, même momentanément, la personne peut difficilement se déplacer.

Le but du projet de recherche Sweet-Home est de définir, à partir d'une étude d'usage conduite auprès d'utilisateurs finals, les fonctionnalités et l'ergonomie d'un système domotique ubiquitaire et attentif, capable d'interagir naturellement avec l'utilisateur. Cette approche, qui entre dans le domaine de l'intelligence ambiante, sera abordée à travers la mise en place d'un contrôleur intelligent communiquant avec les appareils domotiques par des protocoles réseau standard. En effet, un logement a une durée de vie très longue par rapport à celle des équipements ; les technologies employées doivent donc reposer sur des standards bien établis pour une plus grande interopérabilité, un coût réduit et un développement durable. Le système proposé permettra les avancées suivantes :

– assurer une assistance domotique par une interaction naturelle (commande vocale et tactile) ;
– apporter plus de sécurité par la détection de situations de détresse ou d'effraction.

Mes centres d'intérêt dans le projet se situent au niveau de l'étude d'usage qui a été réalisée, du recueil de corpus multimodaux, de l'identification de mots-clés dans la parole, et, de la prise de décision à partir de données multimodales.

Cirdo-Recherche ANR TECSAN 2010 (Innovation Technologique au Service de la Santé), durée 3 ans. Je suis responsable local de ce projet qui finance un ingénieur doctorant au sein de l'équipe GETALP. Personnel permanent LIG : 36 hommes.mois, personnel temporaire LIG : 36 hommes.mois, aide allouée LIG : 197 962€. Labelisation par le pôle de compétitivité Minalogic.

L'objectif recherché est de mettre au point un "Compagnon Intelligent Réagissant au Doigt et à l'Œil" qui représente un produit de télélien social augmenté et automatisé par l'intégration de services innovants (reconnaissance automatique de la parole, analyse de situations (scènes) dans un environnement complexe non contrôlé) visant à favoriser l'autonomie et la prise en charge, par les aidants, de patients atteints de maladies chroniques ou de la maladie d'Alzheimer ou d'affections apparentées.

De plus, ce projet permettra non seulement la validation de technologies génériques, une évaluation psychologique et ergonomique portant sur les usages des services développés (concernant l'utilité, l'utilisabilité et l'accessibilité, l'acceptation, les aspects éthiques, le modèle économique...), mais aussi des enquêtes critiques sur les connaissances acquises par les professionnels des services à la personne (SAP), qui seront ensuite transférées à l'ensemble du secteur.

Mes centres d'intérêt dans le projet se situent au niveau des services d'aide à la personne, de la reconnaissance de la parole des personnes âgées, de la recherche d'information, et de l'analyse de scènes qui sera nécessaire pour une prise de décision ou la génération d'un compte rendu journalier.

4 Activités scientifiques

4.1 Participation à des jurys de thèse

Comme examinateur

– Lyès HAMOUDI (GIP, Ecole des Mines de Douai), soutenue le 10 juin 2011.

Comme encadrant
- Anthony Fleury (TIMC-IMAG), soutenue le 24 octobre 2008.

4.2 Membre de comités scientifiques et relecture d'articles

Relecture d'articles pour des revues internationales
- Medical Engineering and Physics (Elsevier) 2006
- Pattern Recognition Letters (Elsevier) 2008, 2009
- IEEE Transactions on Systems Man Cybernetics 2010, 2011

Comité de relecture de conférences internationales
- Sped (Speech Technology and Human Computer Dialog) 2005, 2007, 2009, 2011
- EMBC 2009, 2010, 2011
- Human 2007

Comité de relecture de conférences nationales
- MajecSTIC 2009

4.3 Enseignement

Tutorat de projets de fin d'étude en école d'ingénieurs (ENSERG puis POLYTECH-TICE)

4.4 Fonctions collectives

Participation au GDR STIC Santé, groupe "e-santé", 2010, 2011.
Participation aux séances de *Brainstorming* du pôle Minalogic, 2008-2011.
Participation aux séances de travail et de réflexion de l'ADEBAG, 2009, 2010.
Rédacteur pour le projet PLUME du CNRS (Promouvoir les Logiciels Utiles, Maitrisés et Economiques pour la communauté de l'Enseignement Supérieur et de la Recherche).

4.5 Collaborations internationales (hors projets)

Institut Polytechnique de Bucarest (Human-Computer Dialogue Group) : échanges scientifiques avec le Pr Corneliu Burileanu, coencadrement d'étudiants de MASTER. Classification des sons de la vie courante dans un habitat intelligent.

Institut Polytechnique de Hanoi (laboratoire MICA) : échanges scientifiques, coencadrement d'étudiants. Reconnaissance de la parole et des sons de la vie courante, espaces perceptifs.

5 Publications

Le tableau ci-dessous fournit une vue synthétique des types de publications par thème. Les abréviations utilisées pour les thèmes sont : TSAP pour Traitement des signaux audio et de parole, ADM pour Analyse de données multimodales et O pour Autres. Pour les types de publications, j'utilise les abréviations suivantes : CL pour chapitre dans des ouvrages internationaux avec comité de relecture, RI pour articles dans des revues internationales avec comité de relecture, RN pour articles dans des revues nationales avec comité de relecture, CI pour articles dans des conférences internationales avec comité de relecture, CN pour articles dans des conférences nationales avec comité de relecture, et WI pour articles dans des colloques internationaux avec comité de relecture.

La liste de publications ne contient aucun article entre 1995 et 2002 car mon activité d'ingénieur de recherche a été consacrée durant cette période à des tâches de service jusqu'en 2000 puis à une reconversion thématique entre 2000 et 2002.

	RI	RN	CL	CI, WI	CN	Total
TSAP	3	2	1	22		28
ADM	2			11	3	16
O	3	1		4	1	9
Total	8	3	1	37	4	53

5.1 Mémoires

[1] "Etude du comportement non linéaire de bulles de gaz au voisinage de leur résonance dans un champ ultrasonore : application à la détection, à la localisation et au calibrage de bulles stationnaires dans un liquide",

Thèse de doctorat d'ingénieur, INSA de Lyon, 1982, 189 p.

5.2 Revues internationales avec comité de lecture (RI)

[1] "Development of Audio Sensing Technology for Ambient Assisted Living : Applications and Challenges"

M. VACHER, F. PORTET, A. FLEURY, N. NOURY

International Journal of E-Health and Medical Communications, vol. **2**(1), pp. 35-54, January-March 2011.

[2] "Improving Supervised Classification of Activities of Daily Living Using Prior Knowledge"

A. FLEURY, N. NOURY, M. VACHER

International Journal of E-Health and Medical Communications, vol. **2**(1), pp. 17-34, January-March 2011.

[3] "SVM-Based Multi-Modal Classification of Activities of Daily Living in Health Smart Homes : Sensors, Algorithms and First Experimental Results"

A. FLEURY, M. VACHER, N. NOURY

IEEE Transactions on Information Technology in Biomedicine, March 2010, vol. **14**(2), pp. 274 -283.

[4] "Embedded Implementation of Distress Situation Identification through Sound Analysis"

D. ISTRATE, M. VACHER, J.-F. SERIGNAT

The Journal on Information Technology in Healthcare, 2008, vol. **6**(3), pp. 204-211.

[5] "Information Extraction From Sound for Medical Telemonitoring"

D. ISTRATE, E. CASTELLI, M. VACHER, L. BESACIER, J.-F. SERIGNAT

IEEE Transactions on Information Technology in Biomedicine, April 2006, vol. **10**(2), pp. 264-274, ISSN : 1089-7771.

[6] "Study of the magnetic flux line patterns in HTSC-crystals by means of decoration technique and image analysis"

WURSTER K., WEISS F., ROUAULT A., AUDIER M., VACHER M.

Physica C, 1994, pp. 235-240.

[7] "Pentagonal phases as a transient state of the reversible icosahedral-rhombohedral transformation in Al-Fe-Cu"

MENGUY N., AUDIER M., GUYOT P., VACHER M.

Phil. Mag. B, 1993, vol. **68**(5), pp. 595-606.

[8] "Nonlinear Behaviour of Micro-bubbles : Application to their Ultrasonic Detection"

M. VACHER, G. GIMENEZ, R. GOUTTE

Acustica, 1984, vol. **54**, pp. 274-283.

5.3 Revues nationales avec comité de lecture (RN)

[1] "Probabilistic Models for Speech and Sound Analysis"

M. VACHER, D. ISTRATE, J.F. SERIGNAT

Annals of the University of Craiova, Automation, Computers, Electronics and Mechatronics, vol. **4**(31), no. 3, 2007, pp. 129-136.

[2] "Système de télésurveillance sonore pour la détection de situations de détresse"

D. ISTRATE, M. VACHER, J.F. SERIGNAT, L. BESACIER AND E. CASTELLI

ITBM-RBM, May 2006, vol.**27**, issue 2, pp. 35-45, Elsevier.

[3] "Réseaux de vortex dans BiSr2CaCu2Ox"

WEISS F, MIRAMOND C., THOMAS O., ROUAULT A., SENATEUR J.P., AUDIER M., VACHER M., GROULT D., HARDY V., PROVOST J.,RUYTER A., SIMON C.

J. Phys. III France, 1994, vol.**4**, p. 2225.

5.4 Chapitre de livre avec comité de lecture (CL)

[1] "Complete Sound and Speech Recognition System for Health Smart Homes : Application to the Recognition of Activities of Daily Living"

M. VACHER, A.FLEURY, F. PORTET, J.-F. SERIGNAT, N. NOURY

New Developments in Biomedical Engineering, Intech Book, ISBN : 978-953-7619-57-2, Feb. 2010, pp. 645 – 673.

5.5 Conférences et workshops internationaux avec comité de lecture (CI,WI)

[1] "Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions"

B. LECOUTEUX, M. VACHER AND F. PORTET

Interspeech 2011, Florence, Italy, Aug. 2011, pp. 2273-2276.

[2] "The SWEET-HOME Project : Audio Technology in Smart Homes to improve Well-being and Reliance"

M. VACHER, D. ISTRATE, F. PORTET, T. JOUBERT, T. CHEVALIER, S. SMIDTAS, B. MEILLON, B. LECOUTEUX, M. SEHILI, P. CHAHUARA AND S. MÉNIARD

EMBC'11, 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, USA, Aug. 30-Sept. 3, 2011, 4p.

[3] "Distant Speech Recognition for Home Automation : Preliminary Experimental Results in a Smart Home"

B. LECOUTEUX, M. VACHER AND F. PORTET

IEEE SPED 2011, Brasov, Romania, May 2011, pp. 41-50.

[4] "Location of an Inhabitant for Domotic Assistance Through Fusion of Audio and Non-Visual Data"

P. CHAHUARA, F. PORTET AND M. VACHER

Pervasive Health, Dublin, Ireland, May 2011, 4p.

[5] "Fusion of Audio and Temporal Multimodal Data by Spreading Activation for Dweller Localisation in a Smart Home"

P. CHAHUARA, F. PORTET, M. VACHER

Space, Time and Ambient Intelligence, Barcelona, Spain. July 16-22, 2011, 6p.

[6] "A Multimodal Corpus Recorded in a Health Smart Home"

A. FLEURY, M. VACHER, F. PORTET, P. CHAHUARA, N. NOURY

Multimodal Corpora : Advances in Capturing, Coding and Analyzing Multimodality, LREC, 18-21 May 2010, Malta, pp. 99-105.

[7] "Challenges in the Processing of Audio Channels for Ambient Assisted Living"
M. VACHER, F. PORTET, A. FLEURY, N. NOURY
HealthCom, Jul. 2010, Lyon, France, pp. 330-337.

[8] "Introducing Knowledge in the Process of Supervised Classification of Activities of Daily Living in Health Smart Homes"
A. FLEURY, N. NOURY, M. VACHER
HealthCom, Jul. 2010, Lyon, France, pp. 322-329.

[9] "Reconnaissance des sons et de la parole dans un Habitat Intelligent pour la Santé : expérimentations en situation non contrôlée"
M. VACHER, A. FLEURY, F. PORTET, J.-F. SERIGNAT, N. NOURY
Actes du 22ème colloque GRETSI : Traitement du Signal et des Images, Dijon, France, Sep. 2009, 4 p.

[10] "Speech recognition in a smart home : some experiments for telemonitoring"
M. VACHER, A. FLEURY, N. GUIRAND, J.-F. SERIGNAT AND N. NOURY
The 5th IEEE Conference on Speech Technology and Human-Computer Dialogue, From Speech Processing to Language Technology (SpeD 2009), Constanta, Roumanie, Jun. 18-21, 2009, pp. 171-179.

[11] "Determining Useful Sensors for Automatic Recognition of Activities of Daily Living in Health smart home"
F. PORTET, A. FLEURY, M. VACHER, N. NOURY
Intelligent Data Analysis in Biomedicine and Pharmacology, Verona, Italy, Jul. 19, 2009, pp. 63-64.

[12] "Traitement des signaux cinématiques pour la détection et la classification des transferts posturaux : le système ACTIM6D"
A. FLEURY, N. NOURY, M. VACHER
Actes du 22ème colloque GRETSI : Traitement du Signal et des Images, Dijon, France, Sep. 2009, 4 p.

[13] "Supervised Classification of Activities of Daily Living in Health Smart Homes using SVM"
A. FLEURY, N. NOURY, M. VACHER
The 31th IEEE EMBS Annual International Conference, "Engineering the Future of Biomedicine", Minnesota, USA, Sep 2-6, 2009, 4 p.

[14] "A wavelet-based pattern recognition algorithm to classify postural transitions in humans"
A. FLEURY, N. NOURY, AND M. VACHER
17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, Aug 24-28, 2009, pp. 2047-2051.

[15] "Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment"
M. VACHER, A. FLEURY, J.-F. SERIGNAT, N. NOURY, H. GLASSON
Interspeech 2008, Brisbane, Australia, Sep. 22-26, 2008, pp. 496-499.

[16] "Sound and Speech Detection and Classification in a Health Smart Home"
A. FLEURY, N. NOURY, M. VACHER, H. GLASSON, J.-F. SERIGNAT
The 30th IEEE EMBS Annual International Conference, "Personalized Healthcare through Technology", Vancouver, British Columbia, Canada, Aug 20-24, 2008, pp. 4644-4647.

[17] "Data Fusion in Health Smart Home : Primary Individual Evaluation of Two Families of Sensors"
A. FLEURY, M. VACHER, H. GLASSON, J.-F. SERIGNAT AND N. NOURY
The 6th International Conference of the International Society for Gerontechnology, 4-7 juin 2008, 6 p.

[18] "Sound Classification in a Smart Room Environment : an Approach using GMM and HMM Methods"
M. VACHER, J.-F. SERIGNAT, S. CHAILLOL
The 4rd International Conference on Speech Technology and Human-Computer Dialogue, Iasi (Roumanie), pp. 135-146, 10-12 mai 2007.

[19] "Generic Implementation of a Distress Sound Extraction System for Elder Care"
D. ISTRATE, M. VACHER, J.-F. SERIGNAT
28th IEEE EMBS Annual International Conference, Aug 30-Sept 3, 2006, New-York, USA, pp. 3309-3312.

[20] "Speech and Sound Use in Remote Monitoring System for Health Care"
M. VACHER, J.-F. SERIGNAT, S. CHAILLOL, D. ISTRATE, V. POPESCU
Lecture Notes in Computer Science, Artificial Intelligence, Text Speech and Dialogue, vol. 4188/2006, 2006, pp. 711-718.

[21] "First Implementation of a Sound/Speech Remote Monitoring Real-Time System for Home Healthcare"
M. VACHER, P. MENENDEZ-GARCIA, J.F.SERIGNAT, D. ISTRATE
The 6th International Conference IEEE Communications 2006, Bucarest (Roumanie), pp. 111-115, 8-10 juin 2006.

[22] "Détection et classification des sons : application aux sons de la vie courante et à la parole"
D. ISTRATE, M. VACHER, J.-F. SERIGNAT
Actes du 20ème colloque GRETSI : Traitement du Signal et des Images, Louvain-la-Neuve, Belgique, Sep. 2005, vol. 1, pp. 485-488, ISBN : 2-87463-001-2.

[23] "Detection and Speech/Sound Segmentation in a Smart Room Environment"
M. VACHER, D. ISTRATE, J.-F. SERIGNAT, N. GAC
The 3rd International Conference on Speech Technology and Human-Computer Dialogue, Cluj (Roumanie), pp. 37-48, 13-14 mai 2005.

[24] "Sound Processing for Health Smart Home"
D. ISTRATE, M. VACHER, E. CASTELLI, C.-P. NGUYEN
2nd International Conference on Smart homes and health Telematics, Singapour, 15-17 septembre 2004, 8 p.

[25] "Sound Detection through Transient Models using Wavelet Coefficient Trees"
M. VACHER, D. ISTRATE, J.-F. SERIGNAT
Complex Systems Intelligence and Modern Technological Applications, Cherbourg (France), pp. 367-372, 19-22 septembre 2004.

[26] "Multichannel smart sound sensor for perceptive spaces"
D. ISTRATE, M. VACHER, J.-F. SERIGNAT, E. CASTELLI
Complex Systems Intelligence and Modern Technological Applications, Cherbourg (France), pp. 691-696, 19-22 septembre 2004.

[27] "Sound Detection and Classification through Transient Models using Wavelet Coefficient Trees"
M. VACHER, D. ISTRATE AND J.-F. SERIGNAT
12th European Signal Processing Conference EUSIPCO, Vienne (Autriche), pp. 1171-1174, 6-10 septembre 2004.

[28] "Sound Detection and Classification for Medical Telesurvey"
M. VACHER, D. ISTRATE, L. BESACIER, J.F. SERIGNAT, E. CASTELLI
2nd Conference on Biomedical Engineering, Innsbruck (Autriche), pp. 395-398, 16-18 février 2004.

[29] "First Steps in Data Fusion between a Multichannel Audio Acquisition and an Information System for Home Healthcare"
G. VIRONE, D. ISTRATE, M. VACHER, J.F. SERIGNAT, N. NOURY, J. DEMONGEOT
IEEE Engineering in Medicine and Biology Society, Cancun (Mexique), pp. 1364-1367, 17-21 septembre 2003.

[30] "Communication between a Multichannel Audio Acquisition and Information System in a Health Smart Home for Data Fusion"
D. ISTRATE, G. VIRONE, M. VACHER, E. CASTELLI AND J.-F. SERIGNAT
IASTED Internet and Multimedia Systems and Applications, Honolulu (Hawaii, USA), 13-15 août 2003, 6 p.

[31] "Life Sounds Extraction and Classification in Noisy Environment"
M. VACHER, D. ISTRATE, L. BESACIER, JF. SERIGNAT, E. CASTELLI
IASTED Signal and Image Processing 2003, Honolulu (Hawaii, USA), 13-15 août 2003, 6 p.

[32] "Habitat Telemonitoring System based on the Sound Surveillance"
E. CASTELLI, M. VACHER, D. ISTRATE, L. BESACIER, JF. SERIGNAT
International Conference on Information Communication Technologies in Health, Samos (Grèce), pp. 141-146, 11-13 juillet 2003.

[33] "Smart Audio Sensor for Telemedecine"
M. VACHER, D. ISTRATE, L. BESACIER, E. CASTELLI, JF. SERIGNAT
Smart Objects Conference, Grenoble (France), pp. 222-225, 15-17 mai 2003.

[34] "Shear transformation and oblique projection in internal space applied to the generation of atomic structural models for Al-Pd-Mn decagonal approximants"
BERAHA L., AUDIER M., DUNEAU M., VACHER M.
Vth Int. Conf. of Quasicristals, Avignon (France), 22-26 mai 1995.

[35] "The Al-Pd-Pm decagonal phase and T, R, tau2-R approximants"
AUDIER M., DUNEAU M., VACHER M.
Colloque National sur les Quasi-cristaux, INSTN Saclay (France), 1-3 juin 1994.

[36] "Relationships between the quasicristalline and approximant phase in the Al-Pd-Mn system ; an application of the linear phason strain field theory"
AUDIER M., DUNEAU M., VACHER M.
ICPM-94, Babha Atomic Research Center. Bombay (Inde). 9-11 mars 1994.

[37] "Detection of stationary bubbles by the measurement of the harmonic distortion of an ultrasonic wave"
M. VACHER, R. GOUTTE, F. LAKESTANI, M. PERDRIX
Second Mediterranean conference on Medical and Biological Engineering
BIOMED 80, Marseille (France), 15-19 septembre 1980.

5.6 Congrès nationaux avec comité de lecture (CN)

[1] "Localisation d'habitant dans un espace perceptif par réseau dynamique"
P. CHAHUARA AND M. VACHER AND F. PORTET
RTE-AFIA 2011, Représentation et Raisonnement sur le Temps et l'Espace, Chambéry, France, 2011, 10 p.

[2] "Localisation d'habitant dans un environnement perceptif non visuel par propagation d'activation multisource"
P. CHAHUARA, M. VACHER, F. PORTET
Majecstic 2010, Bordeaux, France, Oct. 2010, 8 p.

[3] "Application des SVM à la classification automatique des Activités de la Vie Quotidienne d'une personne à partir des capteurs d'un Habitat Intelligent pour la Santé"
A. FLEURY, N. NOURY, AND M. VACHER
XVIèmes Rencontres de la Société Francophone de Classification, Grenoble, France, 2-4 septembre 2009, pp. 33-36.

[4] "Etude des réseaux de lignes de flux magnétiques dans des mono-cristaux de Bi-2212 irradiés aux ions lourds"

WEISS F., MIRAMOND C., ROUAULT A., SENATEUR J.P., THOMASO O., AUDIER M., VACHER M., GROULT D., HARDY V., PROVOST J., RUYTER A., SIMON C.

III^{ème} Journées d'Etudes des Supra-conducteurs à Haute Température Critique, Caen (France), 16-17 nov. 1993.

5.7 Prix / Distinctions

[1] Prix du meilleur article :

Istrate D., Castelli E., Vacher M., Besacier L., Sérignat J.-F. (2007)
"Information extraction from sound for medical telemonitoring"
IMIA Yearbook 2007, 21 : 72-72, IEEE Trans. Inf. Technol. Biomed. Apr. 2006, 10(2) :264-274.

5.8 Séminaires

[1] "Reconnaissance automatique des sons et de la parole : intérêt pour l'aide au maintien à domicile des personnes dépendantes en perte d'autonomie"

MICHEL VACHER, NICOLAS VUILLERME, FRANÇOIS PORTET, YOHAN PAYAN

Forum 4i, Grenoble, 14 mai 2009.

[2] "Surveillance sonore dans un habitat intelligent"

LAURENT BESACIER & MICHEL VACHER

Journées GESO 2008, Images et Sons : Reconnaissance et Analyse, EPFL, Lausanne, 26 juin 2008.

[3] "Analyse sonore dans un Habitat Intelligent pour la Santé"

MICHEL VACHER

Matinées thématiques du LIG, Saint Martin d'Hères, 4 octobre 2007.

Université de Grenoble
École Doctorale de Mathématiques, des Sciences et Technologies de l'Information et de l'Informatique
Discipline : Informatique et Mathématiques Appliquées

Auteur : Michel VACHER
Titre du mémoire de HDR : Analyse sonore et multimodale dans le domaine de l'assistance à domicile
Date de soutenance : 18 octobre 2011

Résumé La moyenne d'âge de la population des pays industriels augmente régulièrement. Les personnes âgées vivant seules sont de plus en plus nombreuses, soit parce qu'elles préfèrent vivre de manière autonome, soit par manque de place dans les institutions spécialisées. Il faut donc trouver des solutions leur permettant de continuer à rester chez elles de manière confortable et sûre. Les habitats intelligents peuvent constituer une de ces solutions.
Un des plus grands défis dans l'Assistance à la Vie Autonome (AVA) est de concevoir des habitats intelligents pour la santé qui anticipent les besoins de leurs habitants tout en maintenant leur sécurité et leur confort. Il est donc essentiel de faciliter l'interaction avec l'habitat intelligent grâce à des systèmes qui réagissent naturellement aux commandes vocales, en utilisant des microphones et pas des interfaces tactiles.
Ce mémoire définit le concept de maison intelligente et présente quelques projets intéressants. Il précise ensuite de quelle manière l'assistance à domicile peut tirer parti de ce concept en s'appuyant sur l'analyse sonore.
L'acceptabilité d'une interface vocale dans le cadre de l'habitat intelligent a été étudiée grâce à une expérience qui a montré quels étaient les souhaits, les attentes et les craintes des utilisateurs âgés, de leurs familles, et des travailleurs sociaux.
L'analyse audio dans la maison intelligente étant un domaine de recherche encore peu exploré, l'intérêt et la manière d'analyser les informations sonores dans un habitat intelligent sont ensuite abordés par une expérience qui a permis de mettre en évidence les défis et les verrous technologiques qui devront être levés pour pouvoir utiliser les informations sonores en complément des autres modalités, et, dans le cas de la parole, la reconnaissance en conditions d'enregistrement distant.
Une solution pratique mettant en œuvre plusieurs microphones est ensuite présentée. Le but envisagé est la réalisation d'un système de commande vocale mettant l'utilisateur en mesure de piloter son environnement non seulement par les interrupteurs et télécommandes classiques, mais aussi par la voix.
L'intérêt de l'information audio combinée à celle des capteurs domotiques est ensuite mis en évidence au travers d'une analyse multimodale permettant de localiser une personne dans un habitat intelligent ou de déterminer son activité. La localisation est nécessaire, par exemple pour avoir connaissance du contexte dans lequel un ordre domotique a été donné. L'activité peut être utilisée pour observer une évolution des habitudes de la personne pour aider à un diagnostic.
Pour finir, le mémoire présente les perspectives de recherche et les projets à venir de l'auteur. Il est accompagné de la reproduction de 4 communications scientifiques publiées dans des congrès sélectifs à comité de lecture.

Abstract The average age of the population in industrialized countries is steadily increasing. Seniors living alone are more numerous, either because they prefer to live independently, or because of a lack of space in the specialized institutions. We must find solutions allowing them to continue to stay at home comfortably and safely. Smart housings can constitute one of these solutions.
One of the biggest challenges in Ambient Assisted Living (AAL) is to develop smart homes that anticipate the health needs of their inhabitants while maintaining their safety and comfort. It is therefore essential to facilitate interaction with the smart home through systems that respond naturally to voice commands, using microphones but no tactile interfaces.
This thesis defines the concept of smart home and describes some interesting projects. It then explains how home assistance can take advantage of this concept thanks to sound analysis.
The acceptability of a voice interface as part of the intelligent home has been studied through an experiment that showed what are the wishes, expectations and fears of older users, their families, and social workers.
Audio analysis in smart homes is a still unexplored research area, the interest and the methods to analyze sound information in a smart home have been studied here in an experiment which helped to highlight the challenges and technological obstacles to be removed in order to use sound information in addition to other modalities and, in the case of speech, distant speech recognition (ASR in remote recording conditions).
A practical solution using several microphones is then presented. The intended purpose is to achieve a voice control system which will allow the user to control their environment not only by conventional switches and remote control devices, but also by voice.
The advantage of the audio information combined with that of home automation sensors is then revealed through a multimodal analysis making it possible to locate a person in a smart home or to determine their activity. Localization is necessary, for example to determine the context in which a home automation command has been issued. The activity can be used to observe changes in the habits of a person to assist in diagnosis.
Finally, the thesis presents the perspectives of research and future projects of the author. It is accompanied by the reproduction of four scientific papers published in refereed selective conferences.