Reconnaissance des sons et de la parole dans un Habitat Intelligent pour la Santé : expérimentations en situation non contrôlée

Michel Vacher, Anthony Fleury, François Portet, Jean-François Serignat, Norbert Noury

Reconnaissance des sons et de la parole dans un Habitat Intelligent pour la Santé : expérimentations en situation non contrôlée

Michel VACHER[1], Anthony FLEURY[2], François PORTET[1], Jean-François SERIGNAT[1], Norbert NOURY[2]

[1]Laboratoire d'Informatique de Grenoble, UMR CNRS/UJF 5217, équipe GETALP
220 rue de la chimie, BP 53, 38041 Grenoble Cedex 9, France

[2]Laboratoire TIMC-IMAG, UMR CNRS/UJF 5525, équipe AFIRM
Faculté de Médecine de Grenoble, bâtiment Jean Roget, F-38706 La Tronche Cedex, France

Michel.Vacher@imag.fr, Anthony.Fleury@epfl.ch
Francois.Portet@imag.fr, Serignat.Jean-Francois@neuf.fr, Norbert.Noury@insa-lyon.fr

Résumé – L'article présente AUDITHIS, un système complet de reconnaissance des sons et de la parole, et son évaluation en environnement réel non contrôlé dans un appartement. Ce système permet de déclencher des alarmes suite à une détection de mots clés de détresse et apporte des informations audio utilisables pour la reconnaissance des Activités de la Vie Quotidienne (AVQ) d'une personne à domicile. Les expérimentations avec des participants jouant un scénario ont montré un taux de bonnes classifications de 8 classes de sons de 72% ainsi qu'un taux de fausse alarmes de 4% et un taux d'alarmes manquées de 30% pour la détection de mots clés de détresse. Concernant la reconnaissance des activités, des tests sur un ensemble de plus de 11 heures d'enregistrement sur 18 modalités ont été réalisés montrant l'intérêt des informations audio et apportant des perspectives d'amélioration.

Abstract – The paper presents the AUDITHIS system designed for on-line recognition of sounds and speech in smart home. The system generates alarms when distress keywords are recognized in the speech stream and complements the classical sensors in smart home by audio information to recognize Activity of Daily Living (ADL). AUDITHIS has been evaluated in a flat with participant realising a predefined scenario. Sounds processing gave 72% of correct classifications for eight classes of sounds and distress keywords detection led to 4% of false alarms rate and 30% of missed alarms rate. For ADL recognition, 11 hours of 18 sensors from 13 participants have been recorded in a health smart home in Grenoble. The evaluation showed that audio information is an important feature to recognize the human activity.

1 Introduction

En 2050, un tiers de la population française dépassera 65 ans (source INSEE). Ce vieillissement amène à réfléchir sur la manière de maintenir au domicile des personnes âgées seules dans des conditions de vie agréables et sûres. À l'heure actuelle, les institutions d'accueil sont confrontées à un manque de place. De nouvelles solutions, basées sur les technologies de l'information et de la communication, sont en développement dans différents laboratoires de recherche et entreprises. Ces solutions passent par l'installation de centrales permettant de contacter un centre d'appel 24h/24, par la création de capteurs intelligents (par exemple pour détecter des situations à risques telle que la chute), ou par l'installation d'un ensemble de capteurs permettant de suivre l'état de santé des personnes. Pour répondre à cet objectif, nous proposons un système multimodal, comprenant capteurs de présence, contacteurs de porte, capteurs environnementaux et microphones, dont le but est, à terme, d'évaluer l'état de santé du résident par sa capacité à réaliser des activités de la vie quotidienne (AVQ ou ADL : *Activities of Daily Living*)[1]. Dans les projets d'habitats intelligents, les détecteurs de présence, les capteurs cinématiques (contraignants) et les caméras (intrusives) sont souvent utilisés alors que les capteurs sonores peuvent s'intégrer à l'habitat pour des commandes vocales (domotique) [2]. L'originalité de nos travaux est d'avoir réalisé et utilisé un système dans un logement afin de mettre en évidence : (1) les problèmes posés par la détection et la reconnaissance en ligne de sons et de parole, (2) les problèmes posés par l'annulation d'un signal sonore parasite notamment la télévision, et (3) l'apport de l'information audio pour classifier les AVQ.

Après avoir présenté la méthode utilisée, nous détaillerons les expérimentations réalisées et nous discuterons les résultats.

2 Méthode

Le traitement sonore par AUDITHIS (décrit figure 1) consiste à capturer le signal sur chacune des voies d'entrée, le traiter et extraire des informations telles que la catégorie (son de la vie courante ou parole), la localisation du son et sa qualité (origine et Rapport Signal sur Bruit – RSB) et le résultat de reconnaissance (classe de son ou parole, probabilités associées) indispensable au système intelligent qui exploite ces données pour la détection des situations de détresse ou la classification des

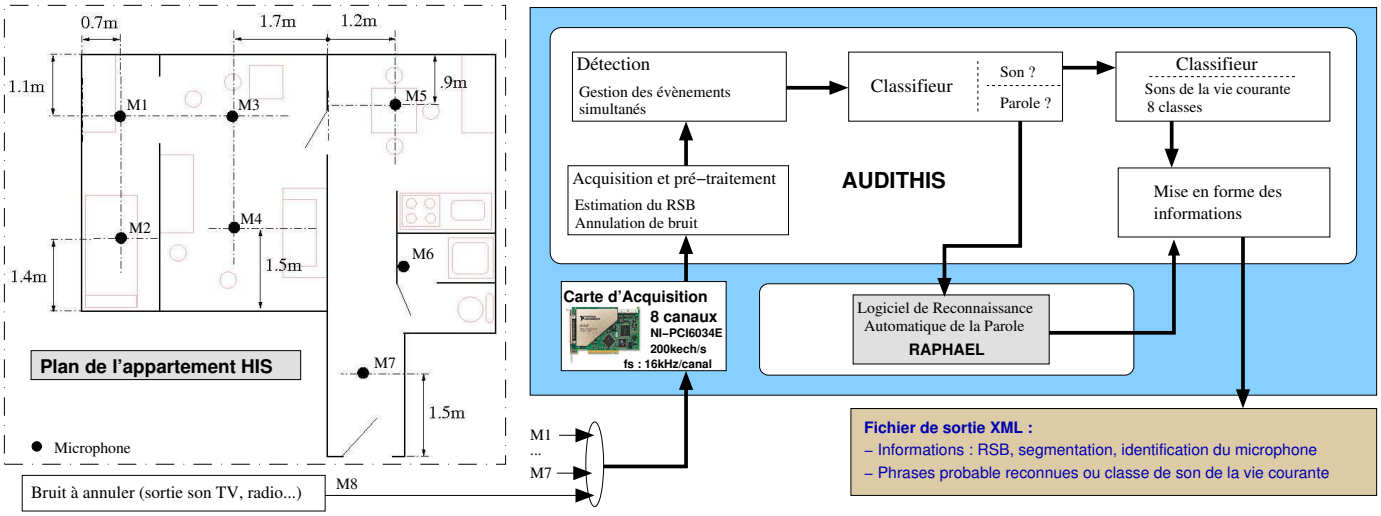

FIG. 1 – Architecture du système AuditHIS et position des microphones dans l'appartement

activités.

La première étape opérée sur chacune des voies est un prétraitement qui fournit une estimation du RSB et permet d'isoler le signal utile dans le bruit de fond sur chacune des voies. La détection se fait par analyse de l'énergie du signal (décomposition en utilisant un arbre d'ondelettes DB6 de profondeur 3) avec seuil adaptatif [3].

La seconde étape est la discrimination entre parole et son qui intervient lorsqu'une détection est complète sur une voie. Elle utilise un classifieur de type Gaussian Mixture Models (GMM). Les paramètres acoustiques utilisés sont des Linear Frequency Cepstral Coefficients (LFCC), la répartition linéaire des bancs de filtre permettant une bonne sensibilité pour les sons de la vie courante. L'apprentissage par Expectation-Maximization (EM) détermine les caractéristiques et le poids des 24 gaussiennes en utilisant des corpus de sons et de paroles enregistrés au laboratoire. Ces corpus comprennent 8 classes de sons de la vie courante (bris de verre, claquement de porte, sonnerie de téléphone, serrure de porte, bruit de pas, bruit de vaisselle, chute d'objet et cris) ainsi qu'un ensemble de paroles prononcées par 21 locuteurs (11 hommes et 10 femmes) afin d'assurer l'indépendance du système vis à vis du locuteur (corpus Anodin/-Détresse) [4]. Le critère BIC (Bayesian Information Criterion) a été utilisé pour déterminer le nombre optimal de gaussiennes. La discrimination son/parole se fait par calcul de maximum de vraisemblance.

Ensuite les sons de la vie courante sont classifiés par un deuxième classifieur en l'une des 8 classes du corpus de sons, à l'aide de modèles GMM ou Hidden Markov Models (HMM) entraînés sur ce corpus. Les performances n'étant pas suffisamment différentes en milieu moyennement bruité (RSB +10 dB), nous utilisons ici des GMM avec 12 gaussiennes apprises sur le corpus de 8 classes de sons en utilisant la même méthode que précédemment et le critère BIC.

La reconnaissance de la parole est effectuée par le système RAPHAEL développé au laboratoire qui fournit les 5 phrases les plus probables pour chaque signal. Les modèles utilisés sont constitués, d'une part des modèles acoustiques multi-locuteur appris sur des grands corpus (300 locuteurs), et, d'autre part des modèles de langage du Français. Le modèle de langage est un système statistique à petit vocabulaire (299 mots) constitué de 299 uni-grammes, 729 bi-grammes et 862 tri-grammes. Il est obtenu à partir d'un corpus de conversation courante de 415 phrases contenant entre autre 39 ordres domotiques et 93 phrases de détresse (« Aïe », « Appelez le SAMU ») dont des phrases syntaxiquement incorrectes (« Ça va pas bien »). Les autres phrases sont des phrases anodines ou se rapportent à une conversation téléphonique (« Allo oui », « À demain », « J'ai bu ma tisane »). Ce corpus inclut les 60 phrases de détresse et les 66 phrases anodines du corpus utilisé pour l'apprentissage du classifieur parole/sons. Le système de reconnaissance n'a pas pour but de reconnaître ce que dit la personne mais uniquement d'extraire des informations utiles concernant une détresse éventuelle ou les activités de la personne (AVQ).

En environnement réel, il peut s'avérer nécessaire d'introduire des traitements supplémentaires avant l'étape de détection pour supprimer des sources sonores parasites comme la télévision ou la radio. À cet effet, il est possible d'introduire un canal de référence afin d'annuler ce signal parasite, avec une méthode d'annulation d'écho. Un estimateur de ce bruit sur chacun des microphones est calculé en utilisant le filtre adaptatif obtenu par l'algorithme Multidelay Block Frequency domain (MDF) [5]. Un post-filtrage du signal suivant la méthode Minimum Mean Square Estimator Short-Time Amplitude Spectrum Estimator (MMSE-STSA) [5] est ensuite utilisé pour réduire l'influence du bruit induit par l'application du filtre adaptatif précédent.

La dernière étape du traitement consiste à utiliser les informations fournies par les capteurs classiques (infrarouges pour la présence, contacteurs de porte ...) et les résultats de l'analyse sonore afin de reconnaître les AVQ telles que dormir, manger, communiquer, etc. Pour étudier l'impact des informations so-

nores sur la reconnaissance d'activités, des méthodes de sélections d'attributs[6] sont appliquées. Il s'agit principalement des méthodes de filtrages (*correlation based* et *consistency based*) et de *wrapping*. De plus, quatre types d'algorithme d'apprentissage ont été appliqués aux ensembles de données issues de la sélection d'attributs. Il s'agit de C4.5 (arbre de décision), de DTM (table de décision), de NBayes (réseau bayésien naïf) et de SVM. Ceux-ci ont été choisis parce qu'ils utilisent des méthodes d'induction très différentes. Cette étape permet d'évaluer l'apport informatif des informations audio ainsi que leur impact sur les méthodes d'apprentissage.

3 Expérimentations et résultats

Pour évaluer un tel système, sept microphones ont été répartis dans un appartement situé à la Faculté de Médecine au laboratoire TIMC-IMAG, ils sont intégrés au plafond. Un autre a été utilisé pour la capture du son du poste TV. La figure 1 présente le système dans sa globalité et montre également la répartition des microphones dans le logement. Cet appartement est aussi équipé de capteurs de présence infrarouges, de contacteurs de porte et d'une station météorologique, totalisant 18 capteurs. Actuellement, la plupart des recherches sont validées par simulation ou dans des environnements très contrôlés et très peu de tests sont effectués en condition réelle. Nous avons donc effectué deux campagnes de mesure dans l'appartement destinées à établir l'intérêt de l'analyse des sons et de la parole concernant la détection de situations de détresse et l'analyse des activités de la vie quotidienne.

3.1 Situation de détresse

La première campagne consistait à effectuer un scénario de 4 actions répétées trois fois en se déplaçant du couloir à la cuisine : (1) claquement d'une porte, (2) faire tomber un objet de la table, (3) crier, (4) faire tourner une cuillère dans une tasse pendant quelques secondes. Ensuite se déplacer dans le séjour et énoncer un ensemble de phrases de détresses et de phrases anodines et enfin répondre 3 fois au téléphone en prononçant un ensemble de 5 phrases correspondant à une conversation téléphonique (par personne : 10 phrases de détresse, 20 anodines et 3 conversations de 5 phrases).

Dix personnes ont participé à cette expérience. Ceci nous a permis de recueillir un corpus de 1 008 sons et de 2 156 phrases avec un RSB moyen de 14,4 $\pm$ 6,5 dB, tous les signaux ayant un RSB inférieur à 5 dB ayant été rejetés. La plupart des évènements sonores ont été captés simultanément par plusieurs microphones avec des RSB différents suivant la position de la personne par rapport au microphone.

La classification des sons donne un taux de bonnes classifications de 72% dans sa globalité. Cependant, certaines disparités sont à noter avec des classes très bien reconnues (sonneries de téléphone à 100%, claquement de porte et chute d'objet à 80%) et d'autres plus difficiles à reconnaître (sons de vaisselle à 51%). Pour cette dernière classe, nous notons que c'est à l'étape de discrimination son/parole qu'intervient le problème car 43% des sons de vaisselle sont classifiés comme de la parole. Ceci peut s'expliquer par le fait que ce son était obtenu en agitant une cuillère dans une tasse en porcelaine vide, le son obtenu était harmonique et donc très proche d'un phonème [a] ou [E]. La classe vaisselle du corpus de son de la vie courante devra être étendue et complétée.

En ce qui concerne la parole, les tests ont été initialement effectués avec un modèle de langage de 9 958 mots construit à partir d'informations textuelles provenant d'Internet et du journal *Le monde*. Ce modèle n'a pas donné de résultats satisfaisants mais cela nous a permis de construire un corpus d'enregistrements en situation sur lequel nous avons pu évaluer les modèles de langage à petit vocabulaire présentés au paragraphe 2.

Pour la suite de l'étude, nous avons utilisé à chaque fois le canal d'enregistrement présentant le meilleur RSB estimé par le système. Par ailleurs, 7 enregistrements saturés ont été supprimés. Cette partie du corpus enregistré par 10 locuteurs comprend alors 429 phrases soit 7,8 minutes de signal de parole, dont 197 phrases de détresse et 232 phrases anodines. Les phrases ont été indexées manuellement car les locuteurs n'ont pas toujours strictement suivi les consignes : certaines phrases n'ont pas été séparées par suffisamment de silence et se trouvent réunies, le contenu a été parfois modifié et la notion de détresse a pu disparaître. Les performances sont évaluées par le TFA (taux de fausses alarmes) et le TAM (taux d'alarmes manquées) :

$$TFA = \frac{nFA}{nPA} \quad , \quad TAM = \frac{nAM}{nPD} \tag{1}$$

nPA désigne le nombre de phrases anodines et nPD celui de phrases de détresse.

Le TFA (mot clé de détresse introduit par le système) et le TAM (mot clé de détresse manqué par le système) sont réduits de moitié en utilisant les modèles de langage à petit vocabulaire, nous obtenons respectivement 4% et 30%.

Par ailleurs, nous avons procédé à une première évaluation de la détection d'appels de détresse en présence d'un signal radio dans un local peu bruité et comportant une pièce unique. Des enregistrements ont été effectués avec 4 participants ayant pour rôle de répéter 2 ou 3 fois 20 phrases de détresse alors que la radio France-Info était diffusée dans la pièce (avec un niveau sonore proche de celui de la parole sur les hauts parleurs). Les performances globales en utilisant l'algorithme d'annulation d'écho donnent un TAM de 27%. Les premiers essais en présence de musique ne sont par contre pas satisfaisants. Ceci peut être attribué au fait que le signal a des composantes sur toute la bande spectrale (percussions notamment).

3.2 Activités de la Vie Quotidienne

Pour évaluer la contribution des informations audio à la reconnaissance d'activités, une seconde campagne d'enregistrement a été menée. Treize personnes (6 femmes, 7 hommes, 30,4 $\pm$ 5,9 ans) ont du réaliser sept activités (AVQ) au moins

une fois, sans contrainte sur la manière, la durée ou l'ordre avec lesquels les activités devaient être effectuées. La durée moyenne des enregistrements était de 51min 40s (23min 11s – 1h 35min 44s, min–max). Les sept activités comprenaient : (1) dormir ; (2) se reposer ; (3) s'habiller ; (4) manger ; (5) aller aux toilettes ; (6) se laver ; et (7) téléphoner. Les 18 capteurs de l'appartement ont permis de dériver 38 attributs numériques et booléens. L'indexation manuelle de cet ensemble de données a été réalisée sur 232 fenêtres de 3 minutes grâce à un ensemble de webcams.

L'application de deux méthodes de sélection d'attributs [6] de type filtre sur cet ensemble a permis de mettre en évidence 11 attributs (sélectionnés plus de 50% de fois lors de la validation croisée stratifiée à 10 tours) majoritairement liés aux capteurs (*PIR — Presence Infra-Red*). Les données décrites par ces attributs constituent l'ensemble **GF** (*Global Filtering*). La méthode wrapper employée avec les algorithmes d'apprentissage C4.5, DTM, NBayes et SVM a permis de mettre en évidence 6 attributs (sélectionnés plus de 50% de fois lors de la validation croisée stratifiée à 10 tours) comprenant 4 attributs liés aux capteurs PIR, 1 attribut lié au son et 1 attribut lié aux contacts de porte. Les données décrites par ces attributs constituent l'ensemble **GW** (*Global Wrapping*). Ces sélections montrent que les capteurs de présence sont les plus informatifs pour la détection d'activités et que les données audio semblent aussi jouer un rôle déterminant. Des apprentissages à partir d'ensemble de données décrits par des sous-ensemble d'attributs ont été menés en utilisant une validation croisée stratifiée à 10 tours et avec 10 répétitions pour évaluer l'impact des attributs sur la détection d'activités. Le tableau 1 résume les résultats obtenus avec la significativité de la différence entre l'ensemble complet et les autres ensembles (t-test).

TAB. 1 – Taux de classifications correctes (%) pour différents algorithmes d'apprentissage et de sélection d'attributs.

méthode	complet	sans audio	PIR seul	GF	GW
C4.5	$83,3$	$76,8^{**}$	$71,7^{**}$	$82,5$	$83,4$
DTM	$80,0$	$75,7^{**}$	$71,3^{**}$	$82,6$	$83,0^{*}$
NBayes	$85,3$	$77,4^{**}$	$72,9^{**}$	$85,1$	$84,5$
SVM	$82,9$	$78,9^{*}$	$75,0^{**}$	$81,3$	$84,6$
moyenne	$82,9$	$77,2$	$72,7$	$82,9$	$83,9$

$^{*}p < 0,05$; $^{**}p < 0,01$

Succinctement, l'information audio joue un rôle primordial pour la classification des activités car toutes les méthodes d'apprentissage ont une baisse de performance très significative lorsque ces informations sont retirées (colonne 'sans audio'). Une ANOVA 3x4 comparant l'ensemble complet GF et GW montre un effet de l'ensemble d'apprentissage presque significatif : F(2,1188) =2,88, p=0,056. Des ANOVAs séparées montrent, par contre, que GW est significativement supérieur à l'ensemble complet (F(1,792)=4,3439, p=0,037) et à GF (F(1,792)=4,2497, p=0,040) alors que ces derniers n'exhibent aucune différence significative entre eux. Cependant, la supériorité de GW est fortement influencée par DTM. Ces tests montrent donc l'intérêt de l'information audio pour la reconnaissance d'activités en ligne. Ces résultats montrent aussi que les méthodes de sélection d'attributs (wrapping) permettent de n'utiliser que 16% des attributs (environ 33% des capteurs) tout en améliorant globalement les performances. Cependant, ces résultats peuvent encore être améliorés si les performances de la reconnaissance des sons augmentent pour amener plus de redondance avec les capteurs de présence (p.ex., un bruit dans un pièce indique une présence).

4 Conclusion

Cet article présente le système AUDITHIS d'analyse des sons et de la parole. Ce système a été utilisé dans un habitat intelligent pour la santé lors de plusieurs expérimentations afin de tester sa robustesse dans un environnement réel. Il a fortement contribué à obtenir un taux de bonnes classifications de 85% dans le cadre de la classification de 7 activités de la vie quotidienne. Pour l'évaluation de situation de détresse par AuditHIS, 72% des sons de la vie courante étaient correctement classifiés et 85% pour la parole. L'amélioration des modèles de langage et des méthodes de rehaussement du signal audio dans le bruit devraient améliorer les résultats sur la parole.

Références

[1] A. FLEURY : *Détection de motifs temporels dans les environements multi-perceptifs – Application à la classification des AVQ d'une personne suivie à domicile par télémédecine.* Thèse de doctorat, Université Joseph Fourier, Grenoble, 2008.

[2] J.-C. WANG, H.-P. LEE, J.-F. WANG et C.-B. LIN : Robust environmental sound recognition for home automation. *IEEE Trans. on Automation Science and Engineering*, 5(1):25–31, Jan. 2008.

[3] M. VACHER, D. ISTRATE et J.-F. SERIGNAT : Sound detection and classification through transient models using wavelet coefficient trees. *In Proceedings of EUSIPCO 2004*, pages 1171–1174, Vienna, Austria, 2004.

[4] M. VACHER, J.F. SERIGNAT, S. CHAILLOL, D. ISTRATE et V. POPESCU : Speech and sound use in a remote monitoring system for health care. LNAI, 4188/2006:711–718, 2006.

[5] J.-M. VALIN et I. B. COLLINGS : A new robust frequency domain echo canceller with closed-loop learning rate adaptation. *In IEEE ICASSP'07, Vol. 1*, pages 93–96, Honolulu, Hawaii, USA, 2007.

[6] Y. SAEYS, I. INZA et P. LARRAÑAGA : A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23: 2507–2517, 2007.