Michel L. Vacher, François Portet

La commande vocale en habitat intelligent : 15 ans d'expérience dans l'équipe GETALP

Volume 4, nº 1 (2023), p. 77-105.

https://doi.org/10.5802/roia.51

La Revue Ouverte d'Intelligence Artificielle est membre du Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 2967-9672

Revue Ouverte d'Intelligence Artificielle
Volume 4, n° 1, 2023, 77-105

La commande vocale en habitat intelligent : 15 ans d'expérience dans l'équipe GETALP

Michel L. Vacher[a], François Portet[a]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France
E-mail : Michel.Vacher@imag.fr, Francois.Portet@imag.fr
URL *:* http://lig-getalp.imag.fr/.

RÉSUMÉ. — La commande vocale suscite actuellement un grand intérêt notamment dans l'habitat intelligent pour l'assistance, la santé et le confort. Depuis 2001, les travaux de l'équipe GETALP dans ce domaine s'appuient sur des allers-retours continus entre collecte de données, recherche, développements d'applications et évaluations expérimentales. Au fil des ans, entre 2001 et 2019, ces travaux ont montré la nécessité de s'attaquer aux problèmes durs de ce domaine d'application tels que l'adaptation en continu à l'utilisateur, la prise en compte de plusieurs locuteurs, la nécessité de fonctionner en ambiance bruitée et l'évaluation en milieu écologique. La démarche de l'équipe a fait usage de nombreuses expérimentations qui ont permis d'enregistrer des corpus mis à la disposition de la communauté. Les travaux et évaluations en habitat intelligent montrent la nécessité d'une approche large en considérant l'acte langagier non seulement comme une information linguistique, mais également comme une information située.

MOTS-CLÉS. — Habitat intelligent, reconnaissance automatique de la parole, compréhension du langage naturel, interaction homme machine.

1. INTRODUCTION

Les *habitats intelligents* se situent au carrefour de la *domotique*, de l'*informatique ubiquitaire* et de l'*intelligence artificielle*. Ils ont pour objectif d'assister leurs habitants dans les situations du quotidien pour améliorer leur qualité de vie. Les premières tentatives de la domotique moderne ont vu le jour au début du 20e siècle avec l'essor de l'électricité, telle la maison électrique de Gëorgia Knap en 1913 [48]. Il faudra cependant attendre les avancées de l'électronique pour voir apparaître une domotique industrielle dans les années 80. Les progrès de l'informatique conjoints avec la miniaturisation et la mise en réseau ont conduit à l'émergence de l'informatique ubiquitaire [83]. La maison intelligente résulte donc de la convergence entre les services offerts par la domotique (automatisation), les capacités de perception et de communication de l'informatique ubiquitaire et de l'aptitude de l'intelligence artificielle à raisonner pour prendre des décisions et interagir avec les habitants. L'habitat intelligent (ou *smart home*) rend donc l'habitat capable d'interagir naturellement avec l'humain, d'automatiser des processus et de proposer des assistances au quotidien.

Les fonctionnalités concernées sont le contrôle centralisé de la maison, le multimédia, la surveillance, la gestion de l'énergie et des appareillages domestiques ou encore la détection de situations à risque [17].

La domotique, bien que très industrialisée, n'a malgré tout pas encore connu de diffusion importante dans le grand public mis à part les systèmes d'alarme ou l'intégration de certains automatismes basiques (par exemple, volets roulants). Elle a surtout intéressé un public technophile et n'a pas pu procurer une qualité de service satisfaisante, car elle a été confrontée à des freins importants : le manque d'interopérabilité des systèmes, leur coût excessif, l'usage de technologies non matures ou une utilité mal définie. Les efforts des grands industriels du domaine ont ainsi été plutôt orientés vers l'équipement de bâtiments industriels (usines, commerces) plus facile à appréhender que l'habitat lui-même dont la durée de vie est sans commune mesure avec celle des technologies.

La maison intelligente reste quant à elle un objectif à atteindre plus qu'une réalisation concrète. Son développement est cependant très actif en raison d'un grand nombre de facteurs : la standardisation des réseaux domotiques, l'arrivée de l'internet des objets et notamment des smartphones, une interopérabilité de plus en plus fiable, les progrès importants réalisés en intelligence artificielle, une meilleure compréhension de l'écologie de l'habitat et, enfin, des besoins mieux définis (énergie, santé, confort).

Les premières réalisations d'habitats intelligents en laboratoire ont vu le jour dès la fin des années 1990 et ont eu pour objectif le suivi de la santé. En effet, le vieillissement rapide de la population[(1)] entraîne une augmentation importante du nombre de personnes dépendantes, elles seraient 1,2 million en 2040 [26]. Celles-ci pourront difficilement être toutes prises en charge par des institutions spécialisées. Le maintien à domicile (MAD) est vu comme une solution viable à ce problème en apportant une assistance en ce qui concerne le confort (compensation du handicap), la sécurité (prévention et détection des situations de danger), la santé (suivi de l'activité et de l'autonomie) ainsi que la communication avec l'entourage et avec l'extérieur qui est essentielle pour la personne isolée à domicile. L'intégration des TIC (Technologies de l'Information et de la Communication) dans l'habitat intelligent pourrait faciliter le confort de vie des personnes âgées et faciliter la gestion de situations problématiques. Dans cette perspective, de nombreux projets académiques consacrés à l'habitat intelligent ont vu le jour aux environs de l'an 2000, ils ont souvent nécessité la réalisation d'un habitat expérimental spécifique tel que *the Neural Network House* [51] à l'université du Colorado ou *House_n* [34] au MIT. Mais à cette époque, la commande vocale n'était pas un objectif et ce n'est que récemment qu'elle a commencée a être présente dans les projets dédiés à l'habitat intelligent.

En France, l'un des premiers habitats intelligents construits en laboratoire est l'Habitat Intelligent pour la Santé (HIS) du laboratoire TIMC-IMAG installé en 1999 dans les locaux de la Faculté de Médecine de Grenoble [62]. Cet appartement de

[(1)] En France, selon les projections de l'INSEE, plus d'un habitant sur quatre aura 65 ans ou plus en 2040, le nombre de centenaires passant de 15 000 en 2010 à 200 000 en 2060 [12]

type 2 pièces/cuisine était équipé de capteurs médicaux (tensiomètre, balance...), de capteurs de présence pour suivre l'activité des personnes et de caméras pour suivre les expérimentations. C'est dans ce contexte que le Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole (GETALP)[(2)] a été l'une des premières équipes de recherche en 2001 à intégrer l'analyse sonore dans l'habitant intelligent dans le cadre d'une collaboration avec le TIMC-IMAG [72]. Le GETALP[(3)] est une équipe pluridisciplinaire (informaticiens, linguistes, phonéticiens, traducteurs et traiteurs de signaux) dont l'objectif est d'aborder tous les aspects théoriques, méthodologiques et pratiques de la communication et du traitement de l'information multilingue (écrite ou orale). La méthodologie de travail du GETALP s'appuie sur des aller-retour continus entre collectes de données, investigations fondamentales, développement de systèmes opérationnels, applications et évaluations expérimentales. Les travaux effectués dans le cadre de l'habitat intelligent sont un exemple phare de l'application de cette méthodologie.

L'idée d'utiliser la modalité auditive comme moyen de perception et d'interaction avec l'habitat est basée sur le fait que l'activité domestique génère un ensemble très riche d'informations acoustiques que cela soit en interaction avec l'environnement (par exemple, bruit de porte), par les appareils de l'habitat (par exemple, machine à laver) ou à travers la voix qui est non seulement un vecteur d'informations linguistiques, mais également paralinguistiques (par exemple, émotion, toux, etc.). L'un des apports les plus évidents de cette modalité est la possibilité de l'interaction vocale. En effet, le langage étant le moyen d'expression le plus naturel, il est particulièrement adapté au grand public (plus facile d'utilisation qu'une interface complexe) et aux situations dites « mains libres ». Par ailleurs, les capteurs acoustiques ont la faculté d'être omnidirectionnels, ce qui dispense les personnes d'être dans un lieu précis pour interagir ou de porter un matériel spécifique [57]. D'une manière générale, la modalité auditive semble être en parfaite adéquation avec la vision de Weiser [83].

Le fait que les grands industriels du GAFAM[(4)] s'intéressent à ce domaine vient conforter ce point de vue. Avec l'émergence des assistants intelligents sur les ordinateurs, les *smart phone* et les *smart speakers*, la commande vocale dans l'habitat n'est pas seulement devenue un domaine de recherche important, mais est également devenue une technologie clé pour l'industrie, comme en témoignent Amazon avec son prix Alexa de plusieurs millions de dollars récompensant les équipes universitaires contribuant à l'amélioration des techniques d'IA conversationnelle[(5)].

Dans le monde de la recherche, on peut également constater un intérêt croissant pour les interfaces vocales en milieu ambiant en commençant par les compagnons intelligents en interaction vocale dyadique [6], les systèmes de dialogue du projet

[(2)] À l'époque il s'agissait en fait de l'équipe GEOD du laboratoire CLIPS qui a fusionné en 2007 avec l'équipe GETA pour fonder l'équipe GETALP du LIG

[(3)] https://lig-getalp.imag.fr/.

[(4)] Apparu au milieu des années 2000, le terme GAFAM est un acronyme formé par la lettre initiale des cinq entreprises Google, Apple, Facebook, Amazon et Microsoft. Plus largement, il inclut en fait la quinzaine d'acteurs d'Internet d'envergure mondiale appelés aussi « les géants de l'internet ».

[(5)] https://developer.amazon.com/alexaprize.

PERS [33] pour la gestion des appels d'urgence. Plus récemment, les projets DIRHA [60] et Voice-Home-2 [11] ont porté leurs efforts sur la reconnaissance de parole distante et sur la réalisation de modèles de reconnaissance de la parole robuste. Si des corpus ont été collectés et mis à disposition de la communauté, ceux-ci sont principalement centrés sur la modalité acoustique et n'intègrent pas de systèmes domotiques industriels. Malgré ces avancées, le développement des interfaces vocales dans l'habitat intelligent doit faire face à un ensemble de défis techniques (bruit acoustique, pluralité des acteurs, émotion, respect de la vie privée, ambiguïté du langage naturel) et méthodologiques (quels usages ? Comment évaluer les systèmes ? Comment intégrer les systèmes dans l'écosystème des habitants ? etc.) qui sont encore très actuels [76]. Le but de cet article est de retracer les évolutions successives du domaine depuis les premières velléités d'interface vocale et de mettre en exergue les défis qu'il reste à relever. Ces évolutions sont retracées à travers les études du GETALP qui ont tout d'abord concerné le suivi d'activité d'un patient (« monitoring ») et la reconnaissance d'appels de détresse et qui sont décrites dans la section 2. Dans une deuxième période, détaillée dans la section 3, la maturité des technologies de la parole et une meilleure compréhension des possibilités qu'offre l'interaction vocale dans l'habitat intelligent ont conduit la recherche vers l'étude de la commande vocale de la domotique, c'est-à-dire la réalisation d'une action en réponse à une parole en prenant en compte le contexte d'interaction. Cette recherche a fait intervenir plusieurs autres laboratoires qui ont pu enrichir la compréhension de l'objet d'étude avec des études d'usage et des études ethnographiques. Enfin, dans une troisième partie, présentée dans la section 4, les progrès effectués jusqu'en 2019 permettent de s'attaquer aux problèmes durs de ce domaine d'application tels que l'adaptation en continu à l'utilisateur, la présence possible de plusieurs locuteurs, la nécessité de fonctionner en ambiance bruitée, l'évaluation en milieu écologique. Au fil des années, la démarche du GETALP pour relever les défis cités a été de « *sortir* » de son domaine académique strict (traitement de la parole) pour adopter une démarche considérant l'ensemble du contexte d'interaction (reconnaissance d'activité), de l'environnement (réseau domotique), de la prise de décision, mais également une approche ayant toujours pour objectif l'humain que cela soit d'un point de vue fonctionnel, ethnographique ou éthique.

2. Utilisation de la modalité auditive dans l'habitat : les prémices

La fin des années 1990 a été caractérisée par un foisonnement d'initiatives visant à développer les outils nécessaires à la mise en place de la télémédecine. Les premières études menées dans l'Habitat Intelligent pour la Santé (HIS) du laboratoire TIMC s'inscrivent dans le cadre de cet effort. Elles avaient essentiellement trait au suivi des personnes fragiles à leur domicile. Dans le cadre du projet AILISA, le HIS a servi de plateforme d'évaluation pour des technologies de télésurveillance médicale et d'assistance en gérontologie afin de les installer dans des chambres hospitalières [53]. Les technologies développées concernaient le suivi de l'état de santé (évolution du poids ou variation du cycle circadien) et pouvaient permettre l'émission d'une alerte en cas d'urgence ou de chute. La figure 2.1 donne le plan de l'appartement HIS. Il était équipé d'appareils et de capteurs fixes (balance, tensiomètre, capteurs de présence

infrarouge...) ou embarqués sur la personne (bouton d'appel, actimètre) reliés par réseau.

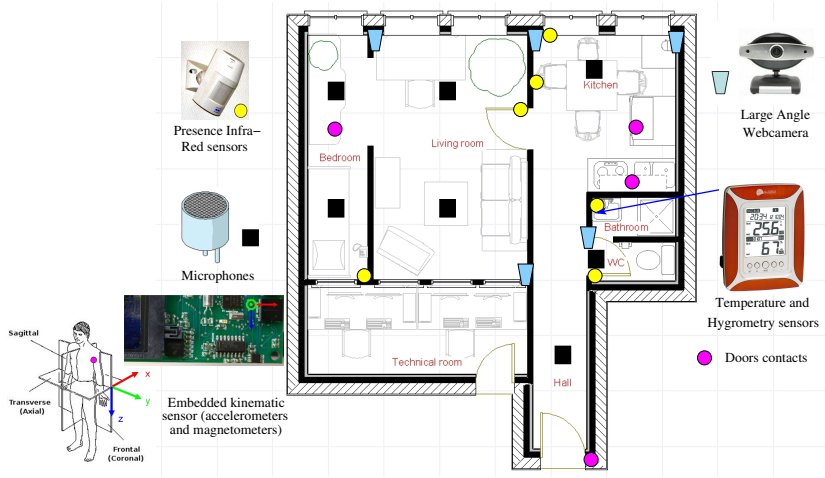


FIGURE 2.1. Positionnement des capteurs dans le HIS (Fleury *et al.*, 2010). L'actimètre est placé sur le corps du participant sous son aisselle gauche (point rouge).

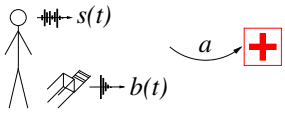
En 2001, l'idée d'équiper l'appartement de microphones permettant une analyse automatique à partir des sons captés dans l'environnement présentait à l'époque une grande originalité, car il n'existait aucune autre initiative dans ce sens. L'objectif annoncé, et mis en œuvre lors du projet RESIDE-HIS [72] financé par la fédération IMAG, répondait aux défis de l'époque : la détection de chutes. En effet, une chute est généralement accompagnée d'un ou de plusieurs sons caractéristiques qui peuvent aider la reconnaissance de ces évènements critiques pour les personnes fragiles. Rapidement, cette analyse a été étendue au cours du projet DESDHIS[(6)] à la reconnaissance des appels à l'aide en s'appuyant sur l'expertise de l'équipe GEOD qui avait développé le système de reconnaissance automatique de la parole RAPHAEL [1, 79] basé sur le système Janus [41]. Enfin, c'est également sur cet élan que s'est développée l'analyse multimodale des Activités de la Vie Quotidienne (AVQ) [39] afin de fournir un monitorage des activités des personnes fragiles isolées pour évaluer leur état d'autonomie en continu. Là encore, l'analyse sonore s'est révélée être une source d'information prometteuse pour cette tâche.

2.1. ANALYSE SONORE DANS L'HABITAT

L'analyse sonore met en jeu à la fois des domaines scientifiques variés comme le traitement du signal, l'apprentissage machine ou l'intelligence artificielle et la reconnaissance automatique de la parole, mais aussi technologiques pour assurer un

[(6)] Le projet DESDHIS (2002-2004) a été financé par une ACI Santé du Ministère de la Recherche

fonctionnement en temps réel. Formellement, à cette époque, le problème pouvait se résumer comme suit.



L'utilisateur de la maison intelligente prononce une commande vocale $s(t)$ qui peut être précédée ou suivie par d'autres évènements sonores $b_i(t)$ (chute de chaise, claquement de porte ...). L'environnement sonore est donc une suite temporelle $X = \{x_j\}$ d'évènements sonores d'intérêt ayant chacun leur intervalle temporel.

L'analyse sonore se définit comme une fonction $f : X \rightarrow a$ qui va extraire chacun des évènements sonores x_j et les analyser. Lorsque x_j est une parole, alors f peut extraire une transcription de l'énoncé $s(t)$ qui peut ensuite être interprétée comme un appel a (par exemple, un appel au secours). Dans le cas contraire, lorsque x_j n'est pas constitué de parole, f peut identifier un type de son, par exemple un bruit caractéristique de chute (nous ne détaillerons pas cette partie dans cet article). Bien entendu, un habitat peut accueillir un ensemble important de microphones, l'analyse doit donc également prendre en compte l'origine des évènements x_j afin de les associer à une localisation sémantique (par exemple, cuisine, salon, etc.).

La première tâche du GETALP a été de construire un système complet et temps réel d'analyse sonore multicanal (chaque pièce est équipée d'au moins un microphone). Le résultat de cet effort est le système AuditHIS [69] dont l'architecture fonctionnelle est décrite sur la figure 2.2.

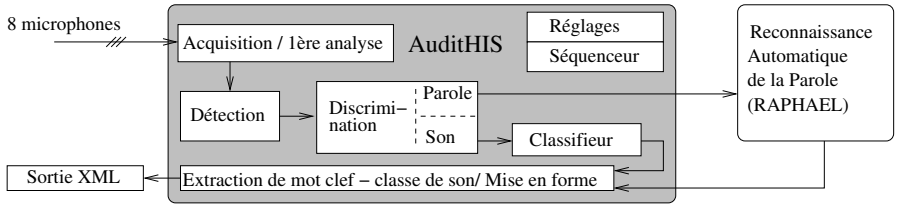


FIGURE 2.2. Organisation des traitements d'analyse sonore en temps réel du logiciel AuditHIS (Vacher *et al.*, 2009).

Le logiciel AuditHIS permet d'acquérir, de détecter et d'analyser les sons et paroles simultanément sur plusieurs microphones. Les modules d'acquisition et de détection décrits figure 2.2 acquièrent et détectent des évènements sonores en permanence indépendamment sur chacune des 8 voies d'entrées. L'hypothèse de travail est que les évènements sonores d'intérêt apparaissent dans l'habitat de manière sporadique au gré des activités de la personne ou du fait de son environnement. Ainsi, ces évènements sont considérés de durée finie et courte, de l'ordre de 1 à 2 secondes. Les bruits de fond, tels que ceux du ventilateur, ne sont au contraire pas détectés par le système, car il s'agit de sons de durée plus longue qui ne sont pas la cible du système. La détection d'évènements sonores est réalisée grâce à une méthode à base d'arbre d'ondelettes qui a été développée par l'équipe [35]. Cette méthode utilise un seuil adaptatif par rapport à l'énergie résiduelle, ce qui permet d'exclure naturellement les bruits de fond non

transitoires. Le rapport signal sur bruit (RSB) de chaque évènement est estimé à cette occasion. Afin d'optimiser le temps calcul, AuditHIS ne traite ensuite que le signal de RSB[(7)] le plus fort après détection sur chaque voie. Le module de discrimination consiste en un classificateur GMM (*Gaussian Mixture Model*) qui permet de différencier entre son et parole, puis lorsqu'il ne s'agit pas d'une parole, un autre classificateur GMM détermine la classe de son parmi 7 classes prédéfinies (claquements de porte, sonneries de téléphone, bruits de pas, bruits de vaisselle, fermetures de serrure, bris de verre, cris). Les évènements identifiés comme étant de la parole sont analysés grâce au système de reconnaissance automatique de la parole RAPHAEL. Enfin, les évènements traités sont ensuite formatés pour être transmis à l'extérieur du logiciel pour traitement ultérieur. Les différents modules de traitement d'AuditHIS ont été entraînés et évalués avec des corpus audio commerciaux ou enregistrés au laboratoire et bruités artificiellement par des enregistrements effectués dans le HIS.

La mise au point d'AuditHIS a permis de rendre la modalité auditive exploitable au même titre que les capteurs domotiques classiques (capteurs infrarouges de mouvement, contacts de porte). Il a ainsi été possible d'utiliser les informations acoustiques pour d'autres tâches que l'appel de détresse ou la chute. La recherche s'est dirigée vers la reconnaissance automatique d'AVQ par classification SVM (Séparateurs à Vaste Marge) à partir de paramètres statistiques calculés sur une fenêtre glissante sur l'ensemble des valeurs des capteurs domotiques.

Pour recueillir les données nécessaires à l'apprentissage du SVM et tester le logiciel AuditHIS en *live*, une expérimentation a été menée en collaboration avec le laboratoire TIMC-IMAG [28]. Treize personnes jeunes et en bonne santé (six femmes et sept hommes) ont joué dans le HIS des scénarios de la vie quotidienne qui comportaient ou non des séquences d'appel à l'aide. Concernant la reconnaissance d'activité, les résultats montraient une exactitude de 86 % avec 7 classes [28], une analyse fine des paramètres a montré l'apport des informations acoustiques [56]. Concernant les résultats de la reconnaissance d'appels de détresse, le taux d'alarmes manquées toujours supérieur à 30 % peut atteindre 85 % pour certains participants, ce qui montre la difficulté de la tâche [71].

2.2. Analyse critique des premiers résultats

Ces années d'études et d'expérimentations ont permis certaines avancées techniques en mettant en oeuvre un système d'analyse sonore temps réel dans un appartement de test réaliste. Les résultats concernant la détection et la discrimination d'évènements sonores se sont révélés être encourageants. Cependant, les conditions réalistes dans lesquelles se sont effectués les tests ont montré la difficulté de la tâche. Les sons recueillis étaient extrêmement divers et pour certains inattendus dans une expérimentation pourtant très contrôlée (pluie, tonnerre, hélicoptère). De nombreux sons enregistrés n'ont pu être identifiés par écoute après les enregistrements, par ailleurs plusieurs sons sont fréquemment mélangés. Dans l'état actuel des connaissances, le

[(7)] Le RSB est calculé sur une fenêtre de 128 ms.

principe de déclenchement d'une alarme à partir de sons reconnus semblait donc illusoire sachant d'autre part que les sons sont fortement bruités (bruit général du bâtiment, musique, circulation routière). Par contre, ce type d'information pourrait s'avérer utile pour compléter les informations provenant des capteurs domotiques.

Les résultats concernant la reconnaissance de la parole se sont révélés très insuffisants. Une première raison tient au niveau de bruit élevé, le rapport signal sur bruit se situant entre 5 et 15 dB. La seconde raison tient aux conditions d'enregistrement très difficiles. Les microphones étaient placés dans chacune des pièces sur le mur. D'une part, la parole enregistrée n'avait pas suivi un trajet direct, mais était réverbérée par plusieurs obstacles, d'autre part, du fait de l'orientation variable des cônes d'émission par la bouche, le rapport entre les hautes et basses fréquences était modifié par la position et l'orientation du locuteur vis-à-vis des microphones. Ces conditions sont celles de la parole distante qui commençait seulement à être étudiée [84], les systèmes de reconnaissance automatique à l'époque se limitant généralement à la dictée vocale, à la reconnaissance de la parole enregistrée dans des émissions radiophoniques ou des réunions avec un microphone placé à proximité et face au locuteur. Il convient d'ajouter à cela la nécessaire adaptation à l'utilisateur, notamment à sa voix. En effet, les voix de personnes âgées, comme celles de personnes souffrant de pathologies particulières (Alzheimer par exemple), sont sous-représentées dans les corpus d'apprentissage et la parole de ces personnes est donc très mal reconnue.

Les expériences ont également révélé la grande dépendance des méthodes nécessitant un apprentissage vis-à-vis des corpus utilisés pour cet apprentissage, ceci concerne les modèles GMM utilisés pour la classification des sons et HMM/GMM pour la reconnaissance de la parole. Les corpus utilisés pour l'entraînement des systèmes se sont révélés inadaptés aux conditions réelles qui sont très hostiles. Par ailleurs, le problème de la détection d'évènements de détresse se heurte à la difficulté méthodologique et éthique de constituer un corpus sonore réaliste et exhaustif de chute sur le sol à cause de l'utilisation obligatoire d'un tapis de protection. Par ces expérimentations et sa volonté de faire progresser la recherche, le GETALP a été l'une des premières équipes à mettre à disposition les corpus collectés[(8)] auprès de la communauté [29].

L'importance de l'expérimentation comme démarche de validation et de découverte de problèmes sous-jacents ou inattendus a également été mise en valeur. Il existe encore des travaux récents focalisés sur la création d'architecture de capteurs sonores répartis et qui n'ont pas encore à ce jour fait l'objet d'évaluations impliquant des données recueillies dans des habitats réels ni de tests en conditions réalistes [52]. Ceci révèle que, si des progrès technologiques importants ont été réalisés lors de ces dernières années, il reste de nombreux efforts à accomplir afin de prendre en charge le milieu particulier de l'habitat ainsi que ses occupants. Nous avons notamment constaté en annotant manuellement les données recueillies, que très souvent la simple écoute ne permettait pas d'identifier un son. Il était alors nécessaire de prendre en compte le contexte dans lequel il avait été émis en utilisant les enregistrements de caméras [76].

[(8)] Disponibles ici : https://lig-getalp.imag.fr/health-smart-home-his-datasets/.

Ceci a donc conduit le GETALP à orienter ses travaux vers une approche multimodale de l'analyse sonore dans l'habitat.

Une autre caractéristique des travaux de recherche menés était leur focalisation sur la résolution technique. Cependant, les expériences pilotes ont fait émerger des questions d'éthique et d'acceptabilité. Il est en effet essentiel que les résultats d'analyse ainsi que les paroles prononcées soient traités en ligne et restent confinés en local, mais aussi que le système soit utilisable et accepté par les utilisateurs. Ceci introduit la nécessité d'intégrer une démarche ethnographique afin de prendre en compte les aspects usages, vie privée et éthique.

Ces résultats montrent que le manque de maturité des techniques d'analyse sonore et de reconnaissance automatique de la parole à cette époque ne permettait pas d'envisager leur déploiement dans quelques appartements pour des tests en condition écologique sur une durée suffisante (plusieurs mois). Ceci est également le résultat de la fin d'une époque où les outils n'étaient pas open source (Janus) et les corpus rares et payants[(9)]. Ceci a conduit à envisager la poursuite de ces études dans le cadre plus général de la commande vocale sensible au contexte conforté par des études ethnographiques et d'usage, avec l'idée de pouvoir apporter ensuite des adaptations pour un public fragile ou suivi médicalement.

3. Expérimentations contrôlées en habitat intelligent : maturité de la reconnaissance automatique de la parole, multimodalité et interactivité

Le début de cette seconde période (2009-2016) a été caractérisé par un regain d'intérêt pour la domotique suite aux progrès de la miniaturisation des équipements électroniques, à la standardisation des bus domotiques (tels que KNX, Zigbee, EnOcean. . .) et à l'apparition de nombreux dispositifs de communication sans fil. L'intelligence des dispositifs était vue comme une simple automatisation de leur fonctionnement, la diffusion auprès du grand public n'a donc pas pris l'essor annoncé à l'époque, essentiellement parce que la qualité des services rendus s'avérait faible. C'est aussi au cours de cette période que les systèmes de Reconnaissance Automatique de la Parole (RAP) ont beaucoup progressé avec la mise à la disposition de la communauté de logiciels libres comme Sphinx [82] puis Kaldi [59]. Alors qu'en 2009, les autres systèmes analysant la parole dans l'habitat étaient très peu nombreux et plutôt orientés vers la réalisation de compagnons assistants [6, 33, 36], de nombreux projets de recherche ont vu le jour vers 2012 s'intéressant à la reconnaissance de la parole bruitée, distante ou adaptative dans l'habitat [21, 22, 27, 32, 80].

C'est à cette période que nos travaux ont pris une dimension résolument multimodale. Ainsi, contrairement aux nombreux travaux spécialisés cités plus haut, notre approche a consisté à considérer le langage naturel situé pour ce qu'il est : plein d'ambiguïtés et d'implicite. C'est pourquoi, bien que la parole soit le vecteur déclencheur

[(9)] On mesure aujourd'hui l'impact du développement de l'open source et de la prolifération des corpus d'apprentissage dans la communauté du traitement automatique de la parole.

d'une commande, la résolution de son intention nécessite d'analyser la situation en cours. Par exemple, lorsqu'une personne en train de dormir dans sa chambre se réveille en pleine nuit et énonce « Allume » il est impossible pour un système purement basé sur le contenu linguistique de prendre une décision. Pourtant, un humain n'aura aucun mal à interpréter l'intention du locuteur. Dans notre approche, l'analyse de la localisation de la personne (chambre), l'activité de la personne (se réveiller), de la période (la nuit), et des dispositifs alentour (lampe de chevet, plafonnier, applique) permet avec un certain degré de certitude de prendre le décision d'allumer la lampe de chevet à faible intensité.

Cette approche a été soutenue par deux projets ANR Sweet-Home [68] et CIRDO [15]. Le projet Sweet-Home, coordonné par le GETALP, visait le public des personnes âgées autonomes utilisant la domotique dans leur vie quotidienne. L'idée était de développer un système domotique que ses utilisateurs installent et utilisent lorsqu'ils sont encore en pleine capacité et qu'il suffise ensuite d'adapter ce système pour assurer la compensation de leur handicap s'ils deviennent fragiles. Ces projets ont été concomitants avec la mise en œuvre de l'appartement DOMUS du LIG [31] dont le plan est donné sur la figure 3.1. Cet appartement réaliste et fonctionnel de 35 m^2 était équipé de plus de 150 capteurs et actionneurs ; outre le bus domotique KNX, plusieurs bus coexistaient, tels que UpnP (Universal Plug and Play) pour la gestion multimédia, X2D pour les contacts de porte et RFID pour l'interaction avec les objets tangibles. Dans le cadre du projet Sweet-Home, chaque pièce de l'appartement a été équipée d'un microphone Sennheiser sans fil placé au plafond et dirigé vers le sol. L'utilisation de microphones sans fil a été privilégiée pour leur facilité d'installation.

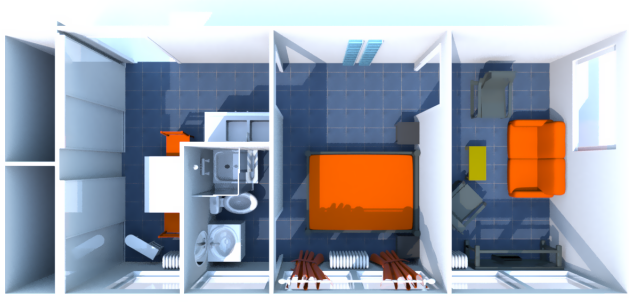


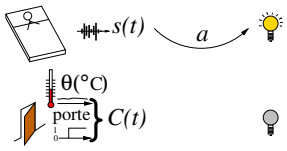
FIGURE 3.1. Vue de dessus de l'appartement intelligent DOMUS de 35 m^2, de gauche à droite : coin repas et cuisine, salle de bains, chambre et bureau.

Dans le projet Sweet-Home, afin d'assurer une conception prenant en compte les utilisateurs cibles, une étude d'usage centrée utilisateur a été conduite en Magicien d'Oz dans DOMUS [57]. Brièvement, les personnes âgées impliquées ont apprécié contrôler la maison par la voix et rejetaient l'utilisation de caméras. Elles pensent que ce système pourrait leur être utile lorsqu'elles auront une vue affaiblie ou des difficultés

à se déplacer. Elles refusent que le système prenne des initiatives à leur place, car cela limiterait leur autonomie. Ceci a confirmé l'intuition que la commande vocale est utile pour assister les personnes dans leur vie quotidienne en les laissant maîtres de leur habitat. Le projet CIRDO a, quant à lui, permis de travailler l'adaptation des méthodes de reconnaissance de la parole aux voix de personnes âgées afin de rendre les interfaces vocales plus performantes pour cette population.

3.1. Vers une commande vocale de la domotique sensible au contexte

Comme cela a été discuté dans la section 2.2, la réalisation d'un système de commande vocale de la domotique nécessite de relever plusieurs défis, et principalement la reconnaissance de la parole en conditions distantes et de la parole spécifique, car les utilisateurs seront souvent âgés ou en situation de détresse lorsqu'ils appellent à l'aide. Par ailleurs, le système doit prendre en compte le contexte pour actionner le périphérique adéquat de la bonne manière. Ainsi, le problème de la commande vocale sensible au contexte pouvait se résumer de la façon suivante.



L'utilisateur de la maison intelligente prononce une commande vocale $s(t)$. L'environnement sonore est une suite temporelle $X = \{x_i\}$ d'évènements sonores d'intérêt ayant chacun leur intervalle temporel. $C = \{c_j\}$ est la suite d'états de l'environnement, c'est-à-dire l'ensemble des valeurs des capteurs à l'instant d'indice j, que nous appellerons contexte.

La commande vocale sensible au contexte se définit comme une fonction $f : (X, C) \rightarrow a$ qui va extraire chacun des évènements sonores x_j, les états de l'environnement (cad le contexte) c_j et les analyser pour donner l'action a à entreprendre. Lorsque x_j est une parole, alors cette parole ($s(t)$) peut être reconnue, puis le contexte c_j est utilisé pour l'interpréter avant d'effectuer une action sur l'environnement.

L'organisation du système de commande vocale est présentée sur la figure 3.2. Le contexte n'est pas directement accessible, il est inféré à partir de données issues de capteurs disposés dans l'environnement et reliés par un réseau domotique. La figure 3.2 (à gauche) illustre la grande diversité de type de données à traiter pour réaliser cette inférence : continues (son), booléennes (état d'un périphérique), évènementielles (capteur de présence infrarouge PIR) ou discontinues (mesure de température à chaque changement). Ce système est articulé autour du séquenceur PATSH [74] illustré 3.2 (à droite) qui a permis l'intégration et l'échange de données en temps réel entre les différents modules (acquisition sonore, différenciation son/parole, Reconnaissance Automatique de la Parole ou RAP, contrôleur intelligent) et leur interaction avec le réseau domotique. L'étage *Acquisition/Détection* limite la durée du signal détecté $s(t)$ à 4,1 s et ne décide de la fin du signal qu'après une période de 0,25 s de silence. Le contrôleur intelligent reçoit de PATSH les résultats de l'analyse sonore, analyse le contexte avant d'émettre un ordre domotique.

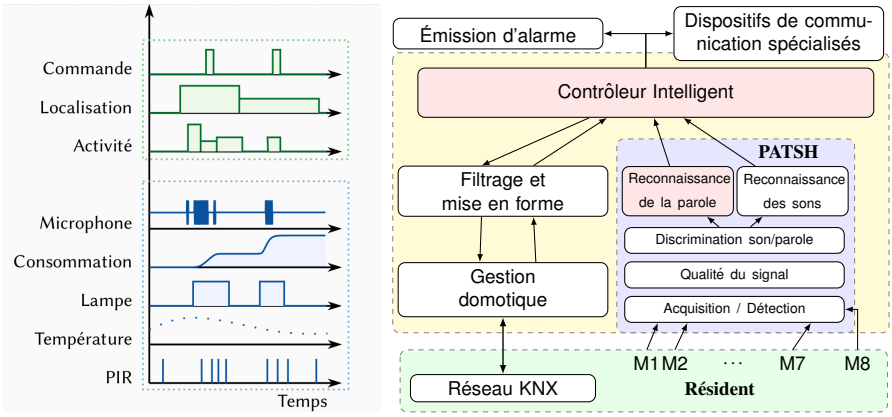


FIGURE 3.2. Système de commande vocale sensible au contexte Sweet-Home. Sur la gauche, la représentation simplifiée des données temporelles que le système doit traiter et inférer. Sur la droite, le diagramme de communication entre systèmes d'analyse sonore PATSH, le contrôleur intelligent et les systèmes extérieurs (bus KNX, systèmes d'alerte et de communication).

3.1.1. *Grammaire des commandes vocales*

Les ordres domotiques ont été définis en utilisant une grammaire très simple qui s'appuie sur nos études préliminaires montrant que les utilisateurs préfèrent des phrases courtes pour piloter leur environnement [57]. Chaque commande est classifiée dans une des catégories suivantes :

- commande de mise en route : (par exemple : « Nestor ferme fenêtre », « Nestor donne l'heure »)
  `mot-clef commande_action objet;`
- commande d'arrêt : (par exemple : « Nestor arrête radio »)
  `mot-clef commande_arrêt [objet];`
- appel à l'aide : (par exemple : « Nestor à l'aide »)
  `[mot-clef] commande_aide`.

Les commandes commencent par un mot-clef permettant de désambiguïser le destinataire de l'énoncé. Son utilisation générale pour tous les systèmes de commande vocale répond à deux objectifs. Tout d'abord, cela permet d'éviter le traitement de phrases n'étant pas des commandes vocales, ce qui respecte le besoin d'intimité des utilisateurs. Ensuite, cela limite l'engorgement du système en limitant les traitements inutiles. Par ailleurs, le choix d'une grammaire simple permet au modèle de langage du système de reconnaissance de la parole d'être plus compact et de réduire le temps de calcul. Dans nos expériences, c'est le mot-clef **Nestor** qui représente l'entité intelligente de l'habitat. Il s'agit d'un choix arbitraire associé facilement au nom d'un majordome pour ceux qui ont lu les albums de Tintin.

3.1.2. *Corpus adapté à la tâche*

L'expérience acquise a montré que, pour développer ce type de système, il est nécessaire de disposer de corpus obtenus dans les conditions de la tâche. Étant donné l'absence de tels corpus à cette époque, nous avons enregistré notre propre corpus dans l'appartement DOMUS. Le protocole suivi pour les enregistrements a bénéficié de l'expérience acquise lors des expérimentations dans le HIS présentées dans la section 2.1. Afin que ce corpus ait un domaine d'utilisation le plus large possible, les traces des capteurs domotiques ont été aussi enregistrées. Les enregistrements ont été répartis en 2 jeux de données.

(1) **Sous-ensemble de données multimodales**. Le but était d'obtenir à la fois des données concernant le contexte dans lequel se trouvait la personne en train d'effectuer des activités de la vie quotidienne seule dans l'appartement (enregistrement des traces du réseau domotique), mais aussi des données concernant la parole en environnement domestique. Les participants restaient libres de la façon dont ils devaient effectuer les activités prévues dans le scénario. En ce qui concerne le son, l'ensemble du flux sonore a été enregistré sur 7 canaux. Les enregistrements des caméras vidéo ont été faits afin de permettre l'annotation du corpus en ce qui concerne la localisation et l'activité de la personne, les sons et les paroles.

    Vingt et une personnes, 7 femmes et 14 hommes entre 22 et 63 ans, ont participé à l'expérimentation sur une durée totale de 26 heures. 1 779 phrases ont été prononcées. L'annotation de ce jeu a été longue surtout en ce qui concerne les sons, car la plupart du temps ils ne peuvent être reconnus par l'annotateur qu'en visualisant la scène grâce aux enregistrements vidéo.

(2) **Sous-ensemble de commandes domotiques**. Ce jeu a été enregistré, car le précédent ne comportait pas de commandes vocales conformément à la grammaire définie dans la section 3.1. Chaque participant seul dans l'appartement devait prononcer des commandes domotiques successivement dans chacune des pièces de l'appartement en les lisant sur une liste, avec, dans certains cas, présence de bruit environnemental (aspirateur, musique ou émission radio).

    Vingt-trois personnes, 9 femmes et 14 hommes entre 19 et 64 ans ont participé, 3 h 6 min de parole ont été enregistrées, soit 5 520 phrases, dont 2 760 bruitées. S'agissant uniquement de parole, l'annotation a été plus facile et rapide.

3.1.3. *RAP en condition distante*

Ce problème était encore peu abordé à l'époque, et afin de résoudre les conséquences acoustiques de la parole distante, le système de RAP a utilisé deux approches conjointes. Premièrement, alors que les modèles acoustiques étaient classiquement appris sur corpus enregistrés en condition studio, nous avons inclus dans nos corpus d'apprentissage des paroles prononcées dans l'appartement dans les conditions de la tâche. Les modèles furent ainsi plus robustes et adaptés aux conditions distantes. Deuxièmement, la commande vocale utilisant un vocabulaire restreint, la modélisation

du langage a été effectuée en interpolant un modèle n-gram générique large vocabulaire (11 018 mots) avec un modèle du domaine appris à partir de la grammaire (88 mots). Ainsi, le décodage était biaisé vers la grammaire de la commande vocale tout en étant capable de reconnaître des termes hors de celle-ci et éviter ainsi les fausses alarmes.

Par ailleurs, la commande vocale dans l'habitat étant résolument multicanal, le système décodait les évènements des deux canaux de plus fort RSB l'un après l'autre. Le décodage d'un canal permettait de pondérer le modèle de langage du système pour le décodage du second en employant une technique appelée *Driven Decoding Algorithm (DDA)* [44, 45]. Ce choix a été fait suite à une étude envisageant les méthodes possibles ayant un faible coût calculatoire.

Pour prendre en compte les erreurs lexicales telles qu'une confusion entre « Nestor » et « les stores », l'identification a été effectuée par mesure d'une distance d'édition (distance de Levenstein) au niveau phonétique entre la meilleure hypothèse de reconnaissance et chacune des commandes possibles dans la grammaire. Si la distance était inférieure à un certain seuil, une commande vocale était détectée, sinon il y avait rejet. Cette approche permet de récupérer certaines erreurs de décodage telles que la déclinaison des verbes ou les variantes de prononciation. En effet, dans de nombreux cas, un mot mal décodé est phonétiquement proche du mot correct.

3.2. Reconnaissance de la voix non standard : voix de personnes âgées

Étant donné que les personnes âgées font partie, avec les personnes malvoyantes, des personnes pour lesquelles l'habitat intelligent serait le plus utile, une étude a été initiée dans l'équipe pour collecter un corpus de voix de personnes âgées. En effet, la RAP durant cette période s'intéressait surtout à la voix audiovisuelle, concentrant principalement ses efforts sur la voix de la population active. La voix de personnes âgées était alors une niche abordée indirectement dans les études considérant la voix pathologique (Alzheimer, dysarthrie) ou de manière exceptionnelle comme dans l'étude [81] dont les auteurs ont utilisé certains enregistrements sonores des plaidoiries de la cour suprême des États-Unis et qui a montré que le taux d'erreur des systèmes croissaient lorsque les orateurs vieillissaient. En 2009, dans le cadre d'une collaboration avec le professeur Alain Franco du service de Gérontologie du CHU de Grenoble, nous avons commencé l'enregistrement d'un corpus de la voix de deux personnes âgées déjà fortement handicapées (82 et 89 ans) ; les séances d'enregistrement étaient fractionnées sur des durées assez courtes afin de ne pas trop fatiguer les locuteurs. Il fut ensuite complété par les enregistrements de 5 personnes autonomes d'âge compris entre 70 et 79 ans à leur domicile. La thèse de Frédéric Aman [2] dans le cadre du projet CIRDO a permis de compléter ce corpus par des enregistrements auxquels ont participé 31 locuteurs jeunes (18 à 59 ans) au laboratoire et 29 locuteurs âgés dans une EPHAD et un centre de résidence pour personnes âgées. Il convient de rappeler que les enregistrements de personnes âgées prennent beaucoup de temps. Le corpus résultant, AD80, est composé des paroles de 95 locuteurs 43 âgés en perte d'autonomie (32 femmes et 11 hommes entre 62 et 94 ans) et 52 non âgés autonomes (27 femmes et 25 hommes entre 18 et 64 ans) [4].

Le corpus AD80 a permis d'établir que, dans le cas de la langue française, une comparaison des scores de vraisemblance de tous les phonèmes pour des groupes de personnes âgées et non âgées montre que les plosives sourdes (p, t, k), les voyelles nasales (ɛ̃, ɑ̃, ɔ̃, œ̃) et les plosives voisées (b, d, g) sont plus affectées par le vieillissement que d'autres phonèmes [5]. Les résultats de ces études indiquent aussi qu'une adaptation des modèles acoustiques à chaque locuteur rapproche les performances de reconnaissance de celles des locuteurs non âgés, ce qui implique toutefois que le système de RAP soit adapté à chaque locuteur. Il subsiste cependant dans ce cas une très forte dispersion des résultats pour la catégorie âgée.

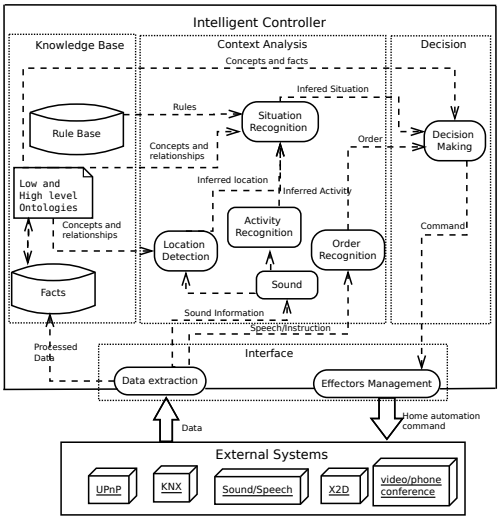
Un autre résultat concerne la mise en évidence d'une corrélation entre le taux d'erreur de mots (*Word Error Rate* – WER) et la dépendance de la personne (critère GIR[10]) mais le trop faible nombre de locuteurs ayant une forte dépendance n'a pas permis de tirer des conclusions plus précises. Il était établi par ailleurs que le déclin des capacités cognitives et perceptives peut fortement dégrader les performances des systèmes de RAP [7, 30]. Cependant, notre étude a permis de montrer, d'une part, que l'âge n'est pas un prédicteur suffisant des performances de la RAP et d'autre part, qu'il est possible d'adapter les systèmes pour obtenir des performances satisfaisantes pour la plupart des individus considérés dans l'étude [66].

3.2.1. *Prise de décision en contexte*

Suite à une commande vocale identifiée, la prise de décision devait se faire en tenant compte du contexte alors que la grammaire ne prévoyait ni d'indiquer le type d'appareil à activer ou à arrêter ni de préciser l'intensité (pour les lumières) ni la pièce concernée. Il était donc nécessaire d'inférer deux informations à partir de toutes les données multimodales disponibles qui caractérisent l'environnement (informations délivrées par les capteurs) : la pièce où se trouve la personne ainsi que l'activité qu'elle réalise au moment de la commande. Dans nos hypothèses, l'appareil actionné est celui de la pièce où se trouve la personne, la lumière étant maximale si la personne lit ou fait la vaisselle, minimale si elle se repose.

La figure 3.3 décrit l'organisation du système de décision (détaillé dans [18]). Brièvement, la connaissance du domaine (configuration des pièces, capteurs, actionneurs) ainsi que les règles régissant les actions sont représentées dans une ontologie. La prise de décision se fait en prenant en compte à la fois la commande reconnue, les données issues de l'ontologie, la situation inférée (activité et localisation). Un ordre est alors transmis à l'actionneur concerné via le réseau domotique. Le modèle de décision est basé sur les réseaux logiques de Markov (MLN – *Markov Logic Network*) [63] afin de modéliser la connaissance experte sous forme de règles logiques pondérées par leur vraisemblance. L'inférence statistique permet donc d'estimer la probabilité de choisir une action, et ce même si les données d'entrées sont incomplètes. Par ailleurs, les poids des règles logiques sont acquis par apprentissage automatique sur corpus. Ce modèle

[10] En France, le degré de dépendance d'une personne âgée est défini par une équipe médico-sociale ou le médecin traitant sur la base de la grille AGGIR (Autonomie, Gérontologie, Groupes Iso-Ressources), qui établit six niveaux de dépendance ou Groupes Iso-Ressources (GIR).



Exemple de scénario joué par le participant dans l'habitat intelligent

Vous allez dans la cuisine
Vous demandez quelle est la température:
Nestor donne la température
Vous prenez une collation
Lorsque vous avez fini, vous posez la vaisselle dans l'évier
Vous demandez l'heure
Nestor donne moi l'heure
Vous réalisez qu'il est tard et vous partez faire vos courses
Avant de quitter l'appartement, vous éteignez les lumières
Nestor éteins la lumière
Vous voulez aussi fermer les rideaux
Nestor baisse les stores
Finalement vous sortez de l'appartement

FIGURE 3.3. Système de prise de décision Sweet-Home [18] et premier scénario utilisé lors des expérimentations dans DOMUS.

appartenant aux modèles statistico-relationels est particulièrement bien adapté à une représentation formelle des connaissances et un raisonnement à partir de capteurs (du monde physique vers le monde de l'automatisme domotique).

3.3. ÉVALUATION EN CONDITIONS RÉALISTES DANS DOMUS

Deux campagnes d'évaluation ont eu lieu dans l'appartement DOMUS (voir figure 3.1) en utilisant le démonstrateur Sweet-Home présenté dans la section 3.1. Les microphones ainsi que les dispositifs domotiques étaient identiques et disposés de la même manière que pour l'enregistrement du corpus (voir description dans la section 3.1.2). Il convient de rappeler que de telles expérimentations nécessitent des efforts importants et que le coût d'enregistrement et d'annotation en résultant est élevé, il a été estimé dans un cas assez similaire à 70 k€ [29].

La première campagne avait pour but la validation fonctionnelle du système. Il a été demandé aux participants de jouer des scénarios de la vie quotidienne et d'interagir avec le système par commande vocale. Les scénarios comportaient aussi une conversation par visioconférence. Seize personnes (7 femmes et 9 hommes) d'âge compris entre 19 et 62 ans ont participé à l'expérimentation sur une durée d'enregistrement totale de 8 h 52 min, ce qui représentait 993 phrases qui ont été annotées.

La deuxième campagne a permis de tirer des conclusions plus pertinentes, car elle a concerné des personnes âgées ou malvoyantes qui sont les utilisateurs potentiels du système [67]. À notre connaissance, il s'agissait de la première expérience de commande vocale en France d'un habitat intelligent utilisant un système complet temps réel au sein de la communauté recherche, le corpus résultant est à ce jour le

seul corpus de ce type disponible pour la recherche. Nous allons donc décrire plus particulièrement cette campagne à laquelle 11 personnes ont participé (6 femmes âgées, 5 personnes malvoyantes, dont 2 femmes). La durée totale du corpus est de 4 h 38 min 59 s, 629 énoncés dont 291 commandes vocales ont été enregistrées. Par rapport à la première campagne, les scénarios avaient été raccourcis et simplifiés afin d'éviter toute fatigue aux participants. Les personnes pouvaient répéter un ordre jusqu'à 3 fois en cas de non-réponse du système, un magicien d'Oz prenant le relai en cas de problème persistant.

Les scénarios 1 et 2 consistaient à effectuer des activités simples nécessitant l'utilisation de commandes vocales, le scénario 1 est présenté sur la figure 3.3. La grammaire n'était pas expliquée et seules les phrases à utiliser étaient fournies. Le scénario 3 était consacré à l'utilisation du système de visioconférence avec un proche ou un ami, puis à la simulation d'appels d'urgence. Le scénario 4 donnait beaucoup plus de liberté à l'utilisateur qui devait générer un certain nombre de commandes vocales sans phrases prédéfinies. Ce scénario a été placé à la fin des autres pour tester si les participants adhéreraient naturellement à la grammaire ou non. Les 4 scénarios nous ont permis de traiter des évènements audio réalistes et représentatifs dans des conditions directement liées aux activités habituelles de la vie quotidienne. L'expérimentation se terminait par un entretien avec un ergonome, l'objectif étant de comprendre le comportement de la personne et d'évaluer l'acceptabilité du système.

D'un point de vue technique, seulement 29 phrases n'ont pas été détectées par le système (4 %) et 85 ont été rejetées (14 %), car, soit leur RSB était trop faible, soit leur durée était en dehors de la plage prévue (en général trop longues). Donc 18 % des phrases n'ont pas été traitées. Il n'y a pas eu de confusion entre les commandes (par exemple ouvrir les stores au lieu des rideaux), mais, pour 41 % des commandes, la sortie du système de RAP ne correspondait pas à une commande valide. Lorsque la radio était en marche suite à une commande, il était souvent impossible de l'arrêter, car aucun dispositif de suppression de bruit n'était prévu. En conséquence des phrases non traitées et du taux d'erreur de mots élevé (43 % pour les commandes vocales), le taux de répétition des commandes a été 76,2 % (93,6 % pour les personnes âgées et seulement 55,4 % pour les déficients visuels). Par ailleurs, les utilisateurs observaient fréquemment une pause entre le mot-clef et le reste de la commande ce qui entraînait une séparation de la commande en 2 énoncés ne respectant pas la grammaire.

Même si la grammaire était très simple, les participants ont eu tendance à en dévier fortement. Une des premières variantes était l'utilisation de l'infinitif au lieu de l'impératif (par exemple, « éteindre la radio » au lieu de « éteins la radio » ou « éteignez la radio »), ce n'est pas naturel, mais cela peut correspondre à la représentation que les personnes se font du langage associé à une machine. Une autre forme consistait en l'utilisation d'une formule de politesse comme « Pouvez-vous éteindre la lumière », « Nestor, éteins la lumière s'il te plaît » ou « S'il te plaît Nestor, éteins la lumière ». Cela tendait à montrer que, pour certains utilisateurs, le système pilotant l'appartement n'est pas seulement un outil, mais qu'il était assimilé à un interlocuteur avec lequel il convient d'interagir avec politesse.

Concernant le système de discrimination parole/non-parole, celui-ci a mal classifié 20 énoncés (dont 5 commandes vocales) qui ont été classifiés comme non-parole, alors que 10 non-paroles ont été classifiées comme parole. Cependant, les expérimentations se sont déroulées dans une ambiance sonore calme mis à part l'utilisation transitoire de la radio. Un autre problème constaté est le temps de réaction du système qui était 1,5 fois la durée d'une commande. Une commande durant en général 1 seconde, il fallait attendre 1,5 seconde de plus pour savoir si elle avait été comprise, temps au cours duquel l'utilisateur restait dans l'incertitude.

3.4. Conclusions sur la période

Au cours de cette période, le GETALP a réalisé un effort important de conception, implémentation et évaluation de systèmes temps réel de commande vocale de la domotique. Les solutions originales étudiées concernent l'ensemble de la chaîne de traitement, l'ensemble des données d'observation de l'habitat (décision sensible au contexte) et prennent en compte les besoins estimés par étude d'usage. Le groupe a également accordé une grande part à l'expérimentation permettant, d'une part de tester la solution technique, et d'autre part de faire émerger des défis scientifiques et une meilleure compréhension de l'objet d'étude. Ces expérimentations ont été valorisées par l'annotation complète et la mise à disposition de corpus [(11)] afin de faire progresser la recherche du domaine [73].

Sur le plan éthique, l'approche défendue par le GETALP est de contraindre les traitements à s'effectuer dans l'habitat afin qu'aucune donnée personnelle ne sorte. C'est une vue à l'opposée des acteurs industriels et de quelques groupes de recherche [21] qui préconisent le traitement déporté sur le « *cloud* » avec les risques que cela représente [42]. De plus, la solution n'utilise aucun capteur vidéo (coûteux, rejeté par les personnes interrogées, sensible à l'éclairage d'ambiance) ni de capteur porté (trop contraignant). Enfin, le système de RAP a été conçu pour n'être performant qu'avec les commandes vocales et non avec les paroles quotidiennes, limitant ainsi l'intrusion dans la vie intime.

Du point de vue de l'usage, les études ont montré que les personnes se méfiaient d'un système prenant les décisions à leur place et qui pourrait les conduire à l'oisiveté. C'est pourquoi nos travaux se sont concentrés sur la commande vocale et non la prédiction d'action. Les personnes ont également trouvé le système trop lent à réagir et se sentaient perdues lors de l'attente. Ceci est dû, d'une part aux traitements qui doivent être améliorés et d'autre part au manque de retour (« *feedback* ») du système. Un autre effet indirect de l'utilisation de la technologie vocale est l'« effet présence » induit par la voix. En effet, les dernières expérimentations incluaient l'usage de la synthèse vocale pour répondre à des questions simples et certaines personnes ont déclaré que le fait de communiquer ainsi ajoutait une présence. L'impact de cet effet reste à mesurer. Enfin, de nombreuses personnes ont dévié de la grammaire et ont affirmé vouloir choisir les termes de celle-ci. L'usage d'une grammaire non figée est donc incontournable,

[(11)] Disponibles ici : http://sweet-home-data.imag.fr/

ce qui impose de comprendre les commandes vocales et rentre dans le cadre de la compréhension du langage naturel (*Natural Language Understanding, NLU*). Enfin, même si ce type de système se destine originellement à une population isolée, un habitat accueille plusieurs personnes, qu'elles soient résidentes (conjoint) ou de passage (petits-enfants, aide ménagère, agent technique, médecin). Les systèmes de commandes vocales doivent donc être capables de fonctionner lorsque plusieurs locuteurs sont présents et d'identifier ceux-ci (par exemple, pour une gestion de droits spécifiques).

Concernant les performances du système Sweet-Home en environnement écologique, l'évaluation dans l'appartement intelligent DOMUS par des utilisateurs potentiels a montré qu'un système ne pourrait pas être utilisé dans le monde réel s'il n'est pas capable de reconnaître la parole en ambiance bruitée et s'il ne s'adapte pas à l'utilisateur et en particulier à son vocabulaire et à la façon dont il souhaite s'adresser au système. Des avancées importantes sont apparues récemment du fait de l'intérêt croissant apporté par la communauté dans ce domaine, intérêt manifesté notamment par l'organisation récurrente des défis CHiME [9, 10] ce qui permet d'envisager d'élargir l'étude au traitement de la parole en environnement bruité.

4. Commande vocale en milieux réalistes : reconnaissance de la parole bruitée et évolution des méthodes de TALN

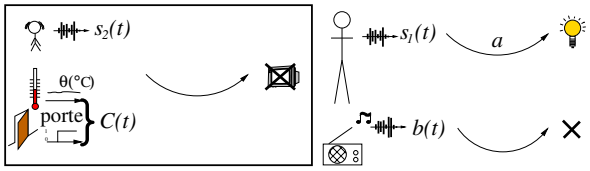
Depuis 2015, la généralisation des réseaux de neurones profonds (*Deep Neural Network*) [64] a apporté un saut de performances important dans la plupart des tâches du traitement du signal et du TALN. Cette émergence est due à la disposition de bases de données massives permettant de tirer parti du pouvoir de généralisation des réseaux de neurones ainsi qu'à la parallélisation des calculs particulièrement efficaces sur les processeurs GPU (*Graphical Process Unit*) optimisés pour le calcul matriciel. Cette évolution a permis au domaine de la séparation de sources et à la modélisation acoustique de changer de paradigme et d'atteindre des performances intéressantes sur des bases de données réalistes [54], d'où le nombre important de travaux concernant la reconnaissance de la parole distante et bruitée [9, 10, 49]. L'industrialisation de la commande vocale est aussi caractéristique de ce début de période que cela soit par les GAFA ou les acteurs des Box internet pour lesquels l'assistance à domicile reliée à une logique commerciale bien particulière n'est pas un but en soi.

Concernant le domaine de la commande vocale, plusieurs problèmes durs semblent donc en passe de trouver une solution satisfaisante : reconnaissance de la parole dans le bruit, identification du locuteur, compréhension du langage naturel, décision. Il n'en reste pas moins que la plupart des études de la littérature ont été conduites hors-ligne sur corpus et que les performances d'une application sous contraintes temps réel dans l'habitat restent inconnues. Ce sont ces objectifs de recherche qui ont motivé le financement par l'ANR du projet VocADom [78] coordonné par le GETALP. Dans le cadre du projet, l'ambition est de parvenir à une commande vocale pouvant être reconnue dans le bruit domestique dans un cadre multi résident et sans contrainte lexicale. Dans cette optique, l'approche consiste de nouveau en une vision large englobant non seulement le signal acoustique, mais également les informations contextuelles tout en

croisant développements techniques, démarche ethnographique et expérimentations. Sur ce dernier point, la volonté du projet est de sortir du laboratoire pour tester la solution développée chez l'habitant.

4.1. La reconnaissance de la parole distante et bruitée multi locuteur

Dans une application domestique réelle, la parole doit être reconnue dans un contexte bruité. Par exemple si un utilisateur prononce une commande vocale $s_1(t)$ pendant l'activité d'un appareil électroménager et pendant qu'un autre résident prononce une commande, $s2(t)$ dans une autre pièce, le problème est redéfini comme suit.



L'environnement sonore et le contexte ont été définis précédemment respectivement comme une suite temporelle $X = \{x_j\}$ d'évènements sonores et la suite d'états de l'environnement $C = \{c_j\}$.

Le contexte c_j est là encore utilisé pour interpréter la commande vocale, $s_1(t)$ ou $s_2(t)$, mais il servira aussi à déterminer la pertinence de la commande. Ainsi, si une personne, telle qu'un enfant, demande à allumer la télévision de manière orale par $s_2(t)$, le système peut, en fonction des droits configurés, décider ou non d'effectuer cette action. Ce type de comportement est également intéressant pour filtrer les commandes pouvant provenir des appareils multimédias (télévision, radio, téléphone).

La reconnaissance de la parole distante et bruitée a fait l'objet d'études spécifiques au travers de projets européens comme DIRHA [22] ou du défi CHiME *Speech Separation and Recognition Challenge* [9] qui est organisé chaque année et utilise maintenant des données enregistrées en conditions réalistes dans des ambiances sonores variées. Le défi est d'adapter les techniques les plus récentes aux conditions propres de l'habitat intelligent, cet aspect est étudié par un des partenaires du projet VocADom [65].

Le rehaussement de la parole dans le bruit nécessite d'utiliser des antennes de microphones avec des espacements connus pour pouvoir calculer des angles d'émission. Ainsi, chacune des pièces des appartements DOMUS et Amiqual4Home du LIG a été équipée de 4 microphones disposés aux angles d'un carré de 10 cm de côté. L'appartement Amiqual4Home possédait un équipement similaire à celui de DOMUS et une superficie de 87 m^2 répartie sur 2 étages [40]. Des corpus a été acquis dans chacun de ces 2 appartements, grâce à des scénarios mettant en jeu simultanément une à quatre personnes ainsi que l'utilisation d'appareils bruyants afin de recueillir des données de plusieurs résidents permettant d'évaluer les algorithmes.

Le premier corpus, VocADom@A4H, a été enregistré dans Amiqual4Home par 11 participants de moins de 25 ans (4 femmes et 7 hommes). Les scénarios de la vie quotidienne comportaient des activités en présence de bruit (aspirateur, musique, radio). Le second corpus, VocADom@DOMUS, a été enregistré dans DOMUS par

12 personnes âgées ou présentant une déficience visuelle. L'enregistrement du second corpus a permis de recueillir, de la part d'utilisateurs cibles (personnes âgées ou ayant une déficience visuelle), des commandes vocales spontanées ainsi que leurs avis concernant l'utilisation du système et aussi les choix du mot-clé. Ces 2 corpus sont multi locuteurs. Pour VocADom@A4H, cet aspect était introduit par la présence d'un expérimentateur rendant visite au participant et dialoguant avec lui. En ce qui concerne VocADom@DOMUS, outre le participant et l'expérimentateur, les scénarios ont été enrichis en ajoutant la visite de deux personnes prétextant un rôle de techniciens d'entretien du système ce qui leur permettait de prononcer des commandes vocales sans perturber le participant. Par ailleurs, une autre personne jouant le rôle d'un stagiaire a été introduite dans le scénario afin que le participant se mette dans le rôle du professeur devant expliquer au stagiaire comment fonctionne le système vocal.

4.2. Compréhension du langage naturel

Les premières expérimentations impliquant des utilisateurs cibles ont montré qu'ils s'écartent très vite d'une grammaire imposée, même très simple, comme celle proposée par le projet Sweet-Home [67]. L'usage d'une grammaire figée n'étant pas envisageable, il faut que le système soit capable de réagir à des commandes définies de manière très flexible. La compréhension du langage naturel (*Natural Language Understanding*, NLU) peut constituer une approche intéressante dans ce cadre.

Nous avons adopté une approche de type *slot-filling* dans laquelle la tâche est, d'une part, d'extraire l'intention principale d'un énoncé et d'autre part, d'étiqueter les attributs/valeurs de l'énoncé. Par exemple dans l'énoncé « Allume la lumière », l'intention est d'agir sur un appareil (`set_device`) tandis que les éléments impliqués sont une action `action=turn_on="allume"` et un appareil `device=light="lumière"` où un attribut est composé de son étiquette (action, device), de ses valeurs normalisées (`turn_on,light`), et du texte associé (`"allume"`, `"lumière"`). Les développements actuels de compréhension utilisent une approche par apprentissage sur données. Les CRF [37] ont été remplacés récemment par des méthodes basées sur les réseaux neuronaux profonds (*Deep Neural Network*, DNN) comme Att-RNN (*Attention-based RNN*) [47]. Alors que la détection d'intention – classification de séquence – a été longtemps considérée comme une tâche séparée de celle d'attribut – étiquetage de séquence – (*slot-filling*), les travaux les plus récents les traitent simultanément étant donné que ces deux tâches sont fortement corrélées. C'est le cas de Tri-CRF qui étend l'étiquetage linéaire de séquence avec un noeud représentant l'acte de dialogue [37], et de Att-RNN qui étend l'encodeur-décodeur RNN d'attribut avec un décodeur d'intention annexe [47].

Attributs et intentions sont fortement liés au domaine d'application, cependant certains travaux visent à une prédiction inter domaine [8] en particulier pour le modèle Tri-CRF [38]. Aucun système n'existant dans le cas de la commande vocale en habitat, il parait très difficile de concevoir un tel système qui soit adapté à des personnes dont les usages et souhaits dans ce domaine sont encore mal connus. Nous avons donc analysé la littérature sur des travaux connexes et utilisé notre expérience pour définir

la sémantique des commandes vocales. Pour cela, nous avons classé les intentions (ou actes de dialogue) en 4 catégories principales :

- **set** : demande de modification de l'état de l'objet ou de l'équipement,
- **check** : demande de vérification de l'état d'un équipement,
- **get** : interrogation sur l'état de l'environnement,
- **contact** : demande d'entrer en communication avec une personne ou un organisme.

Les attributs ont été répartis en 8 catégories : **action** (l'action à effectuer), **device** (l'équipement concerné) **location** (sa localisation), **component** (un composant de l'équipement), **setting** (son réglage), **property** (une propriété de l'environnement, de l'équipement ou de l'endroit) ainsi que **person** et **organisation** (la personne ou l'organisme à contacter). Dans cette étape préliminaire, nous nous sommes limités à la compréhension des commandes contenant une intention unique sans tenir compte du contexte dans lequel elles ont été prononcées ni des commandes précédentes : « s'il te plait je voudrais allumer la lumière » est une commande valide au contraire de « encore un peu » qui se réfère à une commande précédente pouvant concerner un ventilateur par exemple.

Ne disposant pas d'un corpus domotique de grande taille nécessaire aux méthodes de type DNN, nous avons développé un générateur de corpus pour amorcer nos modèles [25]. Le corpus artificiel a été construit à partir d'une grammaire générative utilisant une librairie libre NLTK écrite en Python. Ce corpus contient 28 000 énoncés. De plus, l'expérimentation mentionnée à la section 4.1 a permis de recueillir avec le corpus VocADom@A4H des commandes vocales spontanées puis de plus en plus guidées. Afin d'ajouter en crédibilité, les commandes des participants étaient toujours suivies d'effet grâce à une approche en magicien d'Oz. Ce corpus VocADom@A4H de 4 610 énoncés a été annoté pour tester la compréhension des commandes vocales. Le tableau ci-dessous dresse la comparaison avec le corpus artificiel.

Corpus	Nb. d'intentions	Nb. d'attributs et de valeurs	Nb. d'énoncés
Artificiel	8	17/57	28 000
VocADom@A4H	7	12/46	4 610

Le système Att-RNN a été entraîné sur 90 % du corpus artificiel, les 10 % restant servant au développement. Les tests sur le corpus VocADom@A4H montrent une bonne reconnaissance des intentions (score F1 = 96,70 %), les résultats sont moins bons en ce qui concerne les attributs (score F1 = 74,27 %) et leur valeur (score F1 = 65,05 %). Sur ce même corpus, la méthode Rasa NLU[(12)] à des résultats inférieurs (respectivement 76,57 % 79,03 % 61,95 %) sauf en ce qui concerne les attributs. Cette différence peut s'expliquer par le fait que Rasa NLU utilise des *word embeddings* appris sur un gros corpus indépendant pour extraire les attributs ce qui lui permet d'être plus

[(12)]https://rasa.com/docs/rasa/nlu-training-data/.

robuste aux variations lexicales. Une comparaison détaillée des différentes méthodes envisagées est présentée dans [25, 50].

La composition du corpus artificiel semble être la cause des difficultés à obtenir des résultats satisfaisants. En effet, le corpus VocADom@A4H [55] contient de fortes variations syntaxiques et lexicales. Les répétitions, disfluences et interjections (par exemple « euh ») le rendent très différent du corpus artificiel au niveau syntaxique. La perplexité du modèle de langage 3-gram appris sur le corpus artificiel est 58 sur le corpus VocADom@A4H (hors balise <s>), ce qui assez élevé pour ce type de tâche. Le taux de mots hors vocabulaire est lui aussi élevé, 142 mots n'apparaissant pas dans le corpus artificiel. Ces premiers résultats montrent l'ampleur de la tâche à accomplir afin d'interpréter correctement des commandes vocales sans contrainte de grammaire. Une évolution de cette approche est disponible dans [24].

5. Discussion et conclusion : quel futur ?

Les travaux de l'équipe GETALP sur la commande vocale ont pu être conduits parce qu'il a été possible d'accèder à des appartements intelligents équipés : tout d'abord l'Habitat Intelligent pour la Santé (HIS) du laboratoire TIMC-IMAG puis, dans une seconde période les appartements DOMUS et Amiqual4Home du LIG. Nous avons ainsi pu enregistrer des corpus vocaux et multimodaux et réaliser des expérimentations dans un environnement réaliste. Les résultats de ces expérimentations ont mis en évidence la plupart des défis à relever avant d'envisager une expérimentation en milieu réel, c'est-à-dire au domicile de l'utilisateur. Cependant, les expérimentations faites et les corpus enregistrés comportent un biais. En effet, la durée moyenne de l'expérimentation pour un participant est de 1 h 30 min, elle ne permet pas d'obtenir des données longitudinales caractéristiques du comportement sur une journée complète et encore moins sur un mois.

Nos travaux montrent aussi qu'il est essentiel de disposer de corpus adaptés au domaine d'application, mais que la constitution de corpus et la mise en place d'expérimentations s'avèrent lourdes et coûteuses en temps, y compris en ce qui concerne l'annotation des données enregistrées ; le recrutement de participants d'âge avancé ou présentant un handicap est difficile et le nombre de personnes recrutées peut rarement dépasser la dizaine. Par ailleurs, dans certains cas, comme celui de la compréhension des ordres vocaux, les corpus enregistrés dans un habitat intelligent se sont révélés de taille insuffisante pour couvrir le domaine et nous avons dû recourir à la création de corpus artificiel. Cette création peut certes s'appuyer sur certaines conclusions qui émergent de l'analyse d'expérimentations faites dans des habitats intelligents en magicien d'Oz (corpus Sweet-Home et VocADom) ou avec un démonstrateur autonome (partie du corpus Sweet-Home), mais ces expérimentations restent limitées et ont lieu dans un environnement qui n'est pas le cadre de vie habituel des participants. L'avenir des technologies de traitement du langage naturel repose actuellement sur le *Deep Learning* qui est extrêmement gourmand en données. Une solution que nous avons étudiée consiste à concevoir un système de décision *End-to-End* adaptatif, c'est-à-dire s'adaptant tout au long de la vie du système. Dans ce cadre, les données deviendraient

suffisantes, car fournies toute la journée. Cependant, les premiers résultats que nous avons obtenus montrent des temps d'adaptation encore bien trop longs pour employer un tel système sur le terrain [16]. Comme il n'existe pas à l'heure actuelle de système domotique vocal opérationnel et ouvert, aucune donnée n'a pu être enregistrée au domicile des utilisateurs potentiels. Le défi ne pourra être relevé qu'en développant un démonstrateur susceptible d'être mis en place au domicile de particuliers volontaires afin de réaliser des expérimentations *in vivo*.

L'utilisation par des particuliers à leur domicile met en avant la nécessité de respecter une approche éthique et respectueuse de leur vie privée. La gestion de cet aspect par les grands industriels du GAFA reste obscure. Le paradigme suivi semble être de focaliser toute l'interaction sur les *smart speakers* qui fournissent essentiellement des services accessibles par internet (Question-réponse, musique, achat en ligne). Pour lier l'interaction vocale aux systèmes domotiques, l'approche est d'installer des applications permettant de connecter et configurer les différents systèmes domotiques au *smart speaker* ou tablette ou *smart phone*). Ainsi, les assistants intelligents, en assurant une certaine interopérabilité, sont assurés d'avoir accès à l'ensemble des informations de la maison que les habitants auront bien voulu ouvrir. Bien que les assistants peuvent fonctionner hors ligne dans une certaine mesure, toutes ces informations sont susceptibles d'être transmises et stockées par l'entreprise, voire d'être directement analysées par des employés [23]. Cette démarche, à l'opposé de celle que nous défendons (traitement en local, open source), est celle qui semble être adoptée pour les années qui viennent, les FAI et autres fournisseurs d'IoT allant dans la même direction.

Les années d'expérience du GETALP dans le développement de la commande vocale dans l'habitat permettent de prendre la mesure des évolutions techniques : le développement de l'open source qui contribue à la diffusion rapide des outils, les modèles statistiques puis neuronaux qui ont toujours repoussé les limites des performances, la dépendance grandissante aux corpus de données. Ceci montre que les problèmes réels ne progressent que lorsque l'on considère le problème dans son ensemble et que l'on adopte une approche pluridisciplinaire. Notre objectif de concevoir une solution, utilisable et acceptable par les utilisateurs cibles et la société, ne peut être atteint que si l'utilisateur est partie prenante de la boucle de conception [14] et si l'on prend en considération les effets imprévus (par exemple, illusion d'une présence). Les études futures ne doivent donc pas se limiter à la performance technique, mais s'intéresser aussi aux usages et aux effets des technologies dans l'écosystème de l'habitat domestique.

Remerciements

Les auteurs remercient les membres des différents projets et toutes les personnes qui ont bien voulu consacrer une partie de leur temps pour participer aux expérimentations. L'ANR a soutenu les projets Sweet-Home (ANR-2009-VERS-011), CIRDO (ANR-2010-TECS-012) et VocADom (ANR-16-CE33-0006). François Portet est soutenu par MIAI@Grenoble-Alpes (ANR-19-P3IA-0003)

Bibliographie

[1] M. Akbar & J. Caelen, « Parole et traduction automatique : le module de reconnaissance RAPHAEL », in *Proceedings of COLING-ACL'98* (Montréal, Québec), vol. 2, ACL, 1998, p. 36-40.

[2] F. Aman, « Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile », Thèse, Université de Grenoble, École Doctorale MSTII, 2014.

[3] F. Aman, V. Aubergé & M. Vacher, « Influence of expressive speech on ASR performances: application to elderly assistance in smart home », in *Text, Speech, and Dialogue* (P. Sojka, A. Horak, I. Kopecek & K. Pala, éds.), Lecture Notes in Computer Science, Artificial Intelligence, vol. 9924, Springer International Publishing, Brno , Czech Republic, 2016, p. 522-530.

[4] F. Aman, M. Vacher, S. Rossato & F. Portet, « Analysing the performance of automatic speech recognition for ageing voice: Does it correlate with dependency level? », in *Proceedings of the 4th Workshop SLPAT*, ACL, 2013, p. 9-15.

[5] ———, « In-home detection of distress calls: the case of aged users », in *Proceedings of Interspeech 2013*, ISCA, 2013, p. 2065-2067.

[6] A. Badii & J. Boudy, « CompanionAble – integrated cognitive assistive & domotic companion robotic systems for ability & security », in *Proceedings of SFTAG'09*, SFTAG, 2009, p. 18-20.

[7] L. Baeckman & A. Small, B. aud Whlin, « Aging and memory: cognitive and biological perspectives », in *Handbook of the Psychology of Aging*, 5th ed. Academic Press, San Diego, 2001, p. 349-377.

[8] A. Bapna, G. Tur, D. Hakkani-Tur & L. Heck, « Sequential Dialogue Context Modeling for Spoken Language Understanding », in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue* (Saarbrücken, Germany), ACL, 2017, p. 103-114.

[9] J. Barker, R. Marxer, E. Vincent & S. Watanabe, « The third 'CHIME' speech separation and recognition challenge: Analysis and outcomes », *Computer Speech and Language* **46** (2017), p. 605-626.

[10] J. Barker, S. Watanabe, E. Vincent & J. Trmal, « The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines », in *Proceedings of Interspeech 2018* (Hyderabad, India), ISCA, 2018, p. 1561-1565.

[11] N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina & F. Bimbot, « VoiceHome-2, an extended corpus for multichannel speech processing in real homes », *Speech Commun.* **106** (2019), p. 68-78.

[12] N. Blanpain & O. Chardon, « Projections de population à l'horizon 2060: Un tiers de la population âgé de plus de 60 ans », INSEE (France), 2010.

[13] F. Bloch, V. Gautier, N. Noury, J. L. Lundy, J. Poujaud, Y. E. Claessens & A. S. Rigaud, « Evaluation under real-life conditions of a stand-alone fall detector for the elderly subjects », *Annals of Physical and Rehabilitation Medicine* **54** (2011), p. 391-398.

[14] M.-É. Bobillier-Chaumon, B. Cuvillier, C. Durif-Bruckert, F. Cros, M. Vanhille & S. Bekkadja, « Concevoir une technologie ambiante pour le maintien à domicile : une démarche prospective par la prise en compte des systèmes d'activité », *Le travail humain* **77** (2014), n° 1, p. 39-62.

[15] S. Bouakaz, M. Vacher, M.-E. Bobillier-Chaumon, F. Aman, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, E. Desserée, P. Traineau, J.-P. Vimon & T. Chevalier, « CIRDO: Smart companion for helping elderly to live at home for longer », *Innovation and Research in BioMedical engineering (IRBM)* **35** (2014), n° 2, p. 101-108.

[16] A. Brenon, F. Portet & M. Vacher, « Arcades: A deep model for adaptive decision making in voice controlled smart-home », *Pervasive and Mobile Computing* **49** (2018), p. 92-110.

[17] A. J. B. Brush, B. Lee, R. Mahajan, S. Agarwal, S. Saroiu & C. Dixon, « Home Automation in the Wild: Challenges and Opportunities », in *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, Canada), CHI '11, ACM, 2011, p. 2115-2124.

[18] P. Chahuara, F. Portet & M. Vacher, « Context-aware decision making under uncertainty for voice-based control of smart home », *Expert Systems with Applications* **75** (2017), p. 63-79.

[19] M. Chan, E. Campo, D. Estève & J.-Y. Fourniols, « Smart homes – Current features and future perspectives », *Maturitas* **64** (2009), n° 2, p. 90-97.

[20] Y. Charlon, W. Bourennane & E. Campo, « Mise en œuvre d'une plateforme de suivi de l'actimétrie associée à un système d'identification », in *Symposium Mobilité et Santé (SMS 2011)* (Ax les Thermes (France)), Ludovia, 2011.

[21] H. Christensen, I. Casanueva, S. P. Cunningham, P. D. Green & T. Hain, « HomeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition », in *Proceedings of the 4th Workshop SLPAT*, ACL, 2013, p. 29-34.

[22] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller & P. Maragos, « The DIRHA simulated corpus », in *Proceedings of LREC 2014*, ELRA, 2014, p. 2629-2634.

[23] M. Day, G. Turner & N. Drozdiak, « Amazon workers are listening to what you tell Alexa », 2019, Bloomberg. Consulté le 5 avril 2022, `https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio`.

[24] T. Desot, F. Portet & M. Vacher, « End-to-End Spoken Language Understanding: Performance analyses of a voice command task in a low resource setting », *Computer Speech & Language* **75** (2022), article no. 101369.

[25] T. Desot, S. Raimondo, A. Mishakova, F. Portet & M. Vacher, « Towards a French Smart-Home Voice Command Corpus: Design and NLU Experiments », in *Text, Speech, and Dialogue*, Lecture Notes in Computer Science, Artificial Intelligence, vol. 11107, Springer International Publishing, 2018, p. 509-517.

[26] M. Duée & C. Rebillard, « La dépendance des personnes âgées : une projection en 2040 », *Données sociales – La société française* (2006), p. 613-619.

[27] P. Emanuele, S. Stefano, B. Roberto, F. Giacomo & P. Francesco, « An integrated system for voice command recognition and emergency detection based on audio signals », *Expert Systems with Applications* **42** (2015), n° 13, p. 5668-5683.

[28] A. Fleury, M. Vacher & N. Noury, « SVM-Based Multi-Modal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms and First Experimental Results », *IEEE Transactions on Information Technology in Biomedicine* **14** (2010), n° 2, p. 274 - 283.

[29] A. Fleury, M. Vacher, F. Portet, P. Chahuara & N. Noury, « A French corpus of audio and multimodal interactions in a health smart home », *Journal on Multimodal User Interfaces* **7** (2013), n° 1, p. 93-109.

[30] J. Fozard & S. Gordont-Salant, « Changes in vision and hearing with aging », in *Handbook of the Psychlogy of Aging*, 5th ed. Academic Press, San Diego, USA, 2001, p. 241-266.

[31] M. Gallissot, J. Caelen, F. Jambon & B. Meillon, « Une plate-forme usage pour l'intégration de l'informatique ambiante dans l'habitat : DOMUS », *Technique et Science Informatiques (TSI)* **32** (2013), p. 547-574.

[32] J. F. Gemmeke, B. Ons, N. Tessema, H. Van Hamme, J. Van De Loo, G. De Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. Van Den Broeck, P. Karsmakers & B. Vanrumste, « Self-taught assistive vocal interfaces: an overview of the ALADIN project », in *Proceedings of Interspeech 2013* (Lyon, France), ISCA, 2013, p. 2039-2043.

[33] M. Hamill, V. Young, J. Boger & A. Mihailidis, « Development of an automated speech recognition interface for Personal Emergency Response Systems », *Journal of NeuroEngineering and Rehabilitation* (2009), n° 1, article no. 26 (11 pages).

[34] S. S. Intille, « Designing a home of the future », *IEEE Pervasive Computing* **1** (2002), n° 2, p. 76-82.

[35] D. Istrate, E. Castelli, M. Vacher, L. Besacier & J.-F. Serignat, « Information Extraction From Sound for Medical Telemonitoring », *Information Technology in Biomedicine, IEEE Transactions on* **10(2)** (2006), p. 264-274.

[36] D. Istrate, M. Vacher & J.-F. Serignat, « Embedded Implementation of Distress Situation Identification Through Sound Analysis », *The Journal on Information Technology in Healthcare* **6** (2008), p. 204-211.

[37] M. Jeong & G. G. Lee, « Triangular-Chain Conditional Random Fields », *IEEE Transactions on Audio, Speech, and Language Processing* **16** (2008), n° 7, p. 1287-1302.

[38] M. Jeong & G. G. Lee, « Multi-domain spoken language understanding with transfer learning », *Speech Communication* **51** (2009), n° 5, p. 412-424.

[39] S. Katz, « Assessing Self-maintenance: Activities of Daily Living, Mobility, and Instrumental Activities of Daily Living », *Journal of the American Geriatrics Society* **31** (1983), n° 12, p. 721-727.

[40] P. Lago, F. Lang, C. Roncancio, C. Jiménez-Guarín, R. Mateescu & N. Bonnefond, « The ContextAct@A4H real-life dataset of daily-living activities – Activity recognition using model checking », in *CONTEXT* (Paris, France), LNCS, vol. 10257, Springer, 2017, p. 175-188.

[41] A. Lavie, A. Waibel, L. Levincw, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld & P. Zhan, « Janus-III: speech-to-speech translation in multiple languages », in *Proceedings of ICASSP 97*, vol. 1, IEEE, 1997, p. 99-102.

[42] Le Monde, « Une enceinte connectée d'Amazon envoie une conversation privée par erreur », `https://www.lemonde.fr/pixels/article/2018/05/25/une-enceinte-connectee-d-amazon-envoie-une-conversation-privee-par-erreur_5304453_4408996.html`, 2018, Date : 2018-05-25, Accessed: 2018-09-13.

[43] B. Lecouteux, G. Linarès, J. F. Bonastre & P. Nocéra, « Imperfect Transcript Driven-Speech Recognition », in *Proceedings of InterSpeech'06* (Pittsburg, Pennsylvania, USA), ISCA, 2006, p. 1626-1629.

[44] B. Lecouteux, G. Linarès, Y. Estève & G. Gravier, « Generalized Driven Decoding for Speech Recognition System Combination », in *Proceedings of ICASSP 2008*, IEEE, 2008, p. 1549-1552.

[45] B. Lecouteux, M. Vacher & F. Portet, « Distant Speech Processing for Smart Home Comparison of ASR approaches in distributed microphone network for voice command », *International Journal of Speech Technology* **21** (2018), p. 601-618.

[46] G. Linarès, P. Nocéra, D. Massonié & D. Matrouf, « The LIA speech recognition system: from 10xRT to 1xRT », in *Proceedings of the 10th International Conference on Text, Speech and Dialogue, TSD'07* (Pilsen, Czech Republic), LNCS, vol. 4629, 2007, p. 302-308.

[47] B. Liu & I. Lane, « Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling », in *Proceedings of Interspeech 2016* (San Francisco, USA), ISCA, 2016, p. 685-689.

[48] M. Magnien, « Du rêve à la rigueur : la maison électrique de Georgia Knap », *Culture technique : Machines au foyer* **3** (1981), p. 190-191, Numéro spécial.

[49] M. Malavasi, E. Turri, J. J. Atria, H. Christensen, R. Marxer, L. Desideri, A. Coy, F. Tamburini & P. Green, « An innovative speech-based user interface for smarthomes and IoT solutions to help people with speech and motor disabilities », *Studies in Health Technology and Informatics* **242** (2017), p. 306-313.

[50] A. Mishakova, F. Portet, T. Desot & M. Vacher, « Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in Smart Homes », in *Proceedings of PerDial 2019* (Kyoto, Japan), ISCA/ACL, 2019.

[51] M. C. Mozer, « The neural network house: An environment hat adapts to its inhabitants », in *Proceedings of AAAI Spring Symposium on Intelligent Environments*, vol. 58, 1998, p. 110-114.

[52] J. Navarro, E. Vidaña-Vila, R. M. Alsina-Pagès & M. Hervás, « Real-Time Distributed Architecture for Remote Acoustic Elderly Monitoring in Residential-Scale Ambient Assisted Living Scenarios », *Sensors* **18** (2018), nº 8, article no. 2492, Special Issue: Selected Papers from the 4th International Electronic Conference on Sensors and Applications.

[53] N. Noury et al., « AILISA plateformes d'évaluations pour des technologies de télésurveillance médicale et d'assistance en gérontologie », *Gérontologie et société* **28** (2005), nº 113, p. 97-119.

[54] A. A. Nugraha, A. Liutkus & E. Vincent, « Multichannel Audio Source Separation With Deep Neural Networks », *IEEE/ACM Transactions on Audio, Speech & Language Processing* **24** (2016), nº 9, p. 1652-1664.

[55] F. Portet, S. Caffiau, F. Ringeval, M. Vacher, N. Bonnefond, S. Rossato, B. Lecouteux & T. Desot, « Context-Aware Voice-based Interaction in Smart Home -VocADom@A4H Corpus Collection and Empirical Assessment of its Usefulness », in *PICom 2019 - 17th IEEE International Conference on Pervasive Intelligence and Computing* (Fukuoka, Japan), 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, IEEE, 2019, p. 811-818.

[56] F. Portet, A. Fleury, M. Vacher & N. Noury, « Determining useful sensors for automatic recognition of activities of daily living in health smart home », in *Proceedings of IDAMAP 2009* (Verona, Italy), IMIA, 2009, p. 63-64.

[57] F. Portet, M. Vacher, C. Golanski, C. Roux & B. Meillon, « Design and evaluation of a smart home voice interface for the elderly – Acceptability and objection aspects », *Personal and Ubiquitous Computing* **17** (2013), n° 1, p. 127-144.

[58] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz & S. Thomas, « The subspace Gaussian mixture model—A structured model for speech recognition », *Computer Speech & Language* **25** (2011), n° 2, p. 404-439.

[59] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer & K. Vesely, « The Kaldi Speech Recognition Toolkit », in *Proceedings of IEEE-ASRU* (Hawaii, USA), IEEE SPS, 2011.

[60] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi & M. Omologo, « The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments », in *Proceedings of IEEE-ASRU* (Scottsdale, Arizona, USA), IEEE SPS, 2015, p. 275-282.

[61] V. Rialle, N. Lauvernay, A. Franco, J.-F. Piquard & P. Couturier, « A Smart Room for Hospitalised Elderly People: Essay of Modeling and First Steps of an Experiment », *Technology and Health care* **7** (1999), p. 343-357.

[62] V. Rialle, N. Noury & T. Hervé, « An Experimental Health Smart Home and Its Distributed Internet-based Information and Communication System: First Steps of a Research Project », in *Proceedings of MEDINFO 2001* (London, UK), IOS Press, 2001, p. 1479-1483.

[63] M. Richardson & P. Domingos, « Markov Logic Networks », *Machine Learning* **62** (2006), n° 1-2, p. 107-136.

[64] J. Schmidhuber, « Deep Learning in Neural Networks: An Overview », *Neural Networks* **61** (2015), p. 85-117.

[65] S. Sivasankaran, E. Vincent & D. Fohr, « Keyword Based Speaker Localization: Localizing a Target Speaker in a Multi-speaker Environment », in *Proceedings of Interspeech 2018* (Hyderabad, India), ISCA, 2018, p. 2703-2707.

[66] M. Vacher, F. Aman, S. Rossato, F. Portet & B. Lecouteux, « Making emergency calls more accessible to older adults through a hands-free speech interface in the house », *ACM Transactions on Accessible Computing* **12** (2019), n° 2, article no. 8 (25 pages).

[67] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux & P. Chahuara, « Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation », *ACM Transactions on Accessible Computing* **7** (2015), n° 2, article no. 5 (36 pages).

[68] M. Vacher, P. Chahuara, B. Lecouteux, D. Istrate, F. Portet, T. Joubert, M. E. A. SEHILI, B. Meillon, N. Bonnefond, S. Fabre, C. Roux & S. Caffiau, « The Sweet-Home Project: Audio Technology in Smart Homes to improve Well-being and Reliance », in *Proceedings of EMBC'13* (Osaka, Japan), EMBS, 2013, p. 7298-7301.

[69] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat & N. Noury, « Reconnaissance des sons et de la parole dans un Habitat Intelligent pour la Santé : expérimentations en situation non contrôlée », in *Proceedings of GRETSI 2009* (Dijon, France), 2009, ID456, p. 1-4.

[70] ———, « Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living », in *New Developments in Biomedical Engineering* (D. Campolo, éd.), In-Tech, 2010, p. 645-673.

[71] M. Vacher, A. Fleury, J.-F. Serignat, N. Noury & H. Glasson, « Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment », in *Proceedings of Interspeech 2008* (Brisbane, Australia), ISCA, 2008, p. 496-499.

[72] M. Vacher, D. Istrate, L. Besacier, E. Castelli & J.-F. Serignat, « Smart Audio Sensor for Telemedicine », in *Proceedings of Smart Object Conference (SOC'2003)* (Grenoble, France), Smart Object Conference (SOC'2003), 2003, p. 222-225.

[73] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon & N. Bonnefond, « The Sweet-Home speech and multimodal corpus for home automation interaction », in *Proceedings of LREC 2014* (Reykjavik, Iceland), ELRA, 2014, p. 4499-4506.

[74] M. Vacher, B. Lecouteux, D. Istrate, T. Joubert, F. Portet, M. Sehili & P. Chahuara, « Experimental Evaluation of Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home », in *Proceedings of the 4th Workshop SLPAT*, ACL, 2013, p. 99-105.

[75] M. Vacher, F. Portet, A. Fleury & N. Noury, « Challenges in the Processing of Audio Channels for Ambient Assisted Living », in *IEEE HealthCom 2010 – 12th International Conference on E-health Networking, Application & Services* (Lyon, France), 2010, p. 330-338.

[76] ———, « Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges », *International Journal of E-Health and Medical Communications (IJEHMC)* **2** (2011), n° 1, p. 35-54.

[77] M. Vacher, J.-F. Serignat, S. Chaillol, D. Istrate & V. Popescu, « Speech and Sound Use in a Remote Monitoring System for Health Care », in *Text Speech and Dialogue* (P. Sojka, I. Kopecek & K. Pala, éds.), Speech and Sound Use in a Remote Monitoring System for Health Care, vol. 4188/2006, Springer Berlin/Heidelberg, Brno, Czech Republic, 2006, p. 711 - 718.

[78] M. Vacher, E. Vincent, M.-E. Bobillier Chaumon, T. Joubert, F. Portet, D. Fohr, S. Caffiau & T. Desot, « The VocADom Project: Speech Interaction for Well-being and Reliance Improvement », in *MobileHCI 2018 - 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Barcelona, Spain), 2018.

[79] D. Vaufreydaz, C. Bergamini, J.-F. Serignat, L. Besacier & M. Akbar, « A New Methodology for Speech Corpora Definition from Internet Documents », in *Proceedings of LREC 2000* (Athens, Greece), ELRA, 2000, p. 423-426.

[80] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta & M. Matassoni, « The second CHiME Speech Separation and Recognition Challenge: Datasets, tasks and baselines », in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vancouver, Canada), IEEE, 2013, p. 126-130.

[81] R. Vipperla, S. Renals & J. Frankel, « Longitudinal study of ASR performance on ageing voices », in *Proceedings of Interspeech 2008* (Brisbane, Australia), ISCA, 2008, p. 2550-2553.

[82] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf & J. Woelfel, « Sphinx-4: A Flexible Open Source Framework for Speech Recognition », Tech. report, Sun Microsystems, Inc., Mountain View, CA, USA, 2004.

[83] M. Weiser, « The World is Not a Desktop », *ACM Interactions* **1** (1994), n° 1, p. 7-8.

[84] M. Wölfel & J. McDonough, *Distant Speech Recognition*, John Wiley and Sons, Chichester, UK, 2009, 573 pages.

[85] N. Zouba, F. Bremond, M. Thonnat, A. Anfosso, È. Pascual, P. Malléa, V. Mailland & O. Guerin, « A computer system to monitor older adults at home: Preliminary results », *Gerontechnology* **8** (2009), n° 3, p. 129-139.

Abstract. — Voice control is currently attracting great interest, particularly in smart homes to bring enhanced assistance, health and comfort. Since 2001, the GETALP team's work in this field has been based on continuous back and forth between data collection, research, application development and experimental evaluations. Over the years, since 2001 to 2019, this work has shown the need to address the hard problems of this field of application such as continuous adaptation to the user, the presence of several speakers, the need to operate in a noisy environment, and the evaluation in an ecological environment. The approach of the team uses numerous experiments that made it possible to record corpora which have been made available to the community. The current work and evaluations show the need for a broad approach by considering the language act not only as linguistic information but also as situated information.

Keywords. — Smart home, automatic speech recognition, natural language understanding, Human-computer interaction.

Manuscrit reçu le 2018 septembre 0015, accepté le 8 avril 2022.