Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation

Michel Vacher, Sybille Caffiau, François Portet, Brigitte Meillon, Camille Roux, Elena Elias, Benjamin Lecouteux, Pedro Chahuara

Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation*

MICHEL VACHER, CNRS, LIG, F-38000 Grenoble, France
SYBILLE CAFFIAU, Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
FRANÇOIS PORTET, Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
BRIGITTE MEILLON, CNRS, LIG, F-38000 Grenoble, France
CAMILLE ROUX, Floralis, F-38000 Grenoble, France
ELENA ELIAS, Floralis, F-38000 Grenoble, France
BENJAMIN LECOUTEUX, Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
PEDRO CHAHUARA, Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France

ABSTRACT

This paper presents an experiment with seniors and people with visual impairment in a voice-controlled smart home using the SWEET-HOME system. The experiment shows some weaknesses in automatic speech recognition which must be addressed, as well as the need of better adaptation to the user and the environment. Indeed, users were disturbed by the rigid structure of the grammar and were eager to adapt it to their own preferences. Surprisingly, while no humanoid aspect was introduced in the system, the senior participants were inclined to embody the system. Despite these aspects to improve, the system has been favourably assessed as diminishing most participant fears related to the loss of autonomy.

*Author's addresses: M. Vacher, S. Caffiau, F. Portet, B. Meillon, B. Lecouteux, P. Chahuara, LIG, Campus de Grenoble, 41 rue des Mathématiques, 38041 Grenoble, France; C. Roux, E. Elias, Floralis, 6 allée de Bethléem, 38610 Gières, France.
Reference of the published paper of this draft:
@article{VACHER-TACCESS15,
author = "Michel Vacher and Sybille Caffiau and François Portet and Brigitte Meillon and Camille Roux and Elena Elias and Benjamin Lecouteux and Pedro Chahuara",
title = "Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation",
journal = " ACM Transactions on Accessible Computing (TACCESS) ",
series = "Special Issue on Speech and Language Processing for AT (Part 3)",
volume = "7",
number = "issue 2",
pages = "5:1 - 5:36",
month = "May",
year = "2015",
publisher = "Association for Computing Machinery",
doi ="http://dx.doi.org/10.1145/2738047", }

KEYWORDS: General Terms: Information Systems, Computing Methodologies, Computing Milieux
Additional Key Words and Phrases: assistive technology, voice command, smart home, context-aware inter- action, ambient intelligence, user evaluation
Categories and Subject Descriptors:
I.2.1Artificial IntelligenceApplications and Expert Systems
H.5.2Information Interfaces and PresentationUser Interfaces—natural language
I.2.7Artificial IntelligenceNatural Language Processing—speech recognition and synthesis
K.4.2Computers and SocietySocial Issues—assistive technologies for persons with disabilities .

1 Introduction

Many developed countries are in a demographic transition which will bring the large amount of baby boomers from full-time workers to full-time pensioners. These persons are more likely to live longer than the previous generation but societies have to deal with the rising budgetary costs due to ageing (health and financial support as well as ensuring a good quality of life). In this context, it is likely that families will have to provide more support than in the past century given the reduced availabilities in specialised institutions. Though ageing is challenging the countries' organisation and leads to an increase of disability, this trend stimulates many opportunities to support elderly people in contributing to society. One of the first wishes of this population is to live in their own home as comfortably and safely as possible even if their autonomy decreases. Anticipating and responding to the needs of persons ageing in place is known as Ambient Assisted Living (AAL). In this domain, the development of smart homes is seen as a promising way of achieving in-home daily assistance (**??**). However, given the diverse profiles of the users (e.g., low/high technical skill, disabilities, etc.), complex interfaces should be avoided. Nowadays, one of the best interfaces is the Voice–User Interface (VUI), whose technology is mature and that provides natural language interaction so that the user does not need to learn complex computing procedures (**?**). Moreover, it is well adapted to people with reduced mobility and to some emergency situations (hands-free and distant interaction).

Voice-User Interface in domestic environments has recently gained interest in the speech processing community as exemplified by the rising number of smart home projects that consider Automatic Speech Recognition (ASR) in their design (**??????????**). However, though VUIs are frequent in close domains (e.g., smart phone) there are still important challenges to overcome before implementing VUI in the home (**?**). Indeed, the task imposes several constraints on the speech technology : 1) distant speech condition[1], 2) hands-free interaction, 3) affordable by people, 4) real-time, 5) respect for privacy[2]. Moreover, such technology must be validated in real situations (i.e., real smart homes and users) (**?**).

However, probably one of the main challenges to overcome for successful integration of VUI in AAL is the adaptation of the system to the elderly users. Indeed, the ageing process is characterised by a decay of the main bio-physiological functions, affecting the social role and the integration of the ageing person in the society (**?**). As a consequence, the person withdraws from society to the home and her social and family role are weakened (**?**), but, above all, she losses goals and

[1]Another big challenge is the ability to work in noisy conditions but this is not the focus of this paper, see the CHiME challenge (**?**) for details

[2]Note that as any assistive technology, the intrusiveness of an ICT can be accepted if the benefit is worth it

identity (disengagement theory (**?**)). However, a decrease of activities does not necessarily mean a disengagement in the remaining ones (**?**). Furthermore, through the ageing progression steps, these physiological and mental degradations are irreversible. This progression can be normal, pathological, optimal or successful when the senior succeeds in adapting themselves to the changing situation. Thus, the emergence of disabilities is not originated only in the individual but also in the interaction with his environment. More precisely, it is the unfitness of the environment to the person that causes the incapacity situation (**?**). Overall, elderly people will be less adaptable to a new technology and its limitations (e.g., constraint to pronounce words in a certain way) than younger adults and will present a very diverse set of profiles that makes this population very difficult to design for (**?**). Thus, little is known about the acceptance of VUI in smart homes by elderly people, hence the need for experiments including this population.

In this paper, we present the results of an experiment with seniors and people with visual impairment interacting in a multi-room smart home using voice commands. The inclusion of people with sight impairment is due to the fact that 12% of the population between 65 and 75 years old present an Age-Related Macular Degeneration (ARMD), and due to the demographic change, its prevalence could increase by almost 50% by 2020 (**?**). This experiment required the implementation of a voice-based system, called the SWEET-HOME system. This system uses several mono-microphones set in the ceiling of a smart home equipped with home-automation devices and networks. It selects the "best" sources and employs an ASR decoding and voice command matching which is then analysed by a decision stage that commands the adequate actions to the home automation systems. Hands-free interaction is ensured by constant keyword detection. Indeed, the user must be able to command the environment without having to wear a specific device for physical interaction (e.g., a remote control too far from the user when needed). Despite the fact that setting microphones in a home is a real breach of privacy, by contrast to current smart-phones, we address the problem using an in-home ASR engine rather than a cloud based one (private conversations should not leave the home). Moreover, the limited vocabulary ensures that only the speech utterances relevant to the control of the home are correctly decoded.

The experiment reported in this paper was run in realistic conditions with elderly and people with visual impairment in a fully equipped smart home in order to shed light on the following research questions:

1. Is the ASR performance satisfactory for the application? Does it vary with the user?

2. Is the user able to adapt to the system language?

3. What is the behaviour of the user when interacting with no other feedback than the home automation action?

The analysis of the results was conducted both from quantitative and qualitative points of view on different measures so as to contrast subjective feedback with objective performance. The paper is organised as follows. Section **??** provides background on the VUI processing and the evaluation of smart homes in the AAL context. Section **??** details the technical aspect of the SWEET-HOME system while Section **??** introduces the experimental protocol. The data acquired are summarised in Section **??** and analysed in Section **??**. The results of this analysis are discussed in Section **??** and the paper ends with the conclusion and an outlook of future research directions.

2 Related work

2.1 Voice User Interface in Smart Homes

A rising number of studies about audio technology in smart homes have been conducted which include speech recognition (**?????**), sound recognition (**???**), speech synthesis (**?**) or dialogue (**????**). These systems are either embedded into the home automation system or in a smart companion (mobile or not) or both (see the Companions (**?**) or CompanionAble (**?**) projects for information about this trend). Whatever the system in which they are embedded, VUIs are generally composed of five main components:

1. a Voice Activity Detection (VAD) stage when hands-free interaction is required;
2. an Automatic Speech Recognition (ASR) stage;
3. a Natural Language Understanding (NLU) stage, when required;
4. a decision stage and;
5. a communication stage.

The VAD role is to identify the speech segments in the continuous acoustic stream (i.e., distinguishing the actual speech from the background noise)[3]. Each time a segment is detected, the ASR system outputs the main transcription hypotheses that can then be analysed by the NLU stage. Once a message is inferred, the decision stage chooses the best actions to perform (e.g., changing the temperature) and generates the corresponding commands to send to the home automation system. In case communication with the user is required (e.g., an answer to the question "what is the temperature?"), a communication module translates the answer into natural language. For instance, to generate a natural language response, as might be required in a natural language dialogue system, the communication stage may consist of Natural Language Generation (NLG) and Text-To-Speech (TTS) stages. This processing chain is a general view of the approach but many different approaches exist in the literature (see for instance (**?**) for a more integrated language-free ASR/NLU stage). Due to the focus of this paper on the voice interface, the VAD, ASR and communication stages are detailed below. Indeed, the paper is not about dialogue and is concerned with small vocabulary making the NLU stage marginal. Regarding the decision process, it is detailed in section **??**.

The VAD stage is mainly concerned with the identification of the speech segments by computing spectral features from the signal and applying a thresholding strategy to detect the onset and the end of speech segments. In the particular context of voice command, only speech corresponding to voice commands must be retained and not private conversations. There are two main approaches to handling this aspect:

1. a button-based approach which, when pushed, indicates to the system that the user is about to utter a command or;
2. a keyword sequence which indicates the start of a command.

[3]The VAD can be preceded by a noise reduction stage (see (**?**) for background about the noise reduction challenges in a domestic environment).

The first solution is adopted by most of the commercial systems (e.g., smart phones) but is not hands-free and is not adapted to daily usage (e.g., while washing up) or emergency situations (e.g., far from the button when stuck on the floor after a fall). The second solution, less frequently employed, necessitates keyword detection either at the signal level or language level. It requires a continuous processing of the stream, but this solution makes distant speech possible which thus frees the user from permanently wearing a dedicated device.

The speech recognition is performed by ASR systems which are typically composed of an *acoustic model* (AM) and a *language model* (LM). The acoustic model models acoustic units of speech (generally phones) to be recognized from the acoustic stream. The LM is often represented by n-gram models that represent the successions of words related to the domain. From each frame of the acoustic stream, a vector of spectral features is computed. From this sequence of vectors, the acoustic model, together with a pronunciation dictionary, produces the possible hypotheses about the words that might have been uttered by the speaker. Then the LM considerably reduces the number of possible hypotheses by discarding any sequence of words not in accordance with the domain.

State-of-the-art ASR systems are challenged when applied to VUI applications in smart homes for AAL. One issue is that elderly people have long been ignored by the community when building acoustic models while ageing affects the voice and movement of the person. Indeed, an aged voice is characterized by some specific features such as imprecise production of consonants, tremors, hesitations and slower articulation (**?**). From an anatomical point of view, some studies have shown age-related degeneration with atrophy of vocal cords, calcification of laryngeal cartilages, and changes in muscles of the larynx (**??**). The consequence is that general purpose speaker-independent ASR performance decreases with elderly voices. This phenomenon has been observed in different languages (**???**). Moreover, Vipperla et al. (**?**) used some audio recordings of the proceedings of the Supreme court of the United States of America from the later half of 1990s to 2008 and highlighted a constant degradation of ASR performances for the same person as age increased. In the case of the French language, Aman et al. (**?**) compared the likelihood scores of all the phones for aged and non-aged groups and showed that mid vowels, nasals and phones that depend on the ability to create constrictions with the tongue tip for articulation are more affected by ageing than other phonemes. The results of these studies show that an adaptation of the AM to each speaker makes the ASR performances closer to that with non-aged speakers, however, this implies that the ASR must be adapted to each speaker. These studies were all done for English and French, except another study for European Portuguese which confirmed that the chronological age is a global explanatory factor (**?**). However, this last study also emphasises that many other effects can also be responsible for ASR performance degradation such as decline in cognitive and perceptual abilities (**??**). Moreover, since smart home systems for AAL often concern distress situations, it is unclear whether a distressed voice will challenge the applicability of these systems. Recent studies suggest that ASR performance decreases in case of emotional speech (**?**), however it is still an under-researched area.

Another issue for acoustic processing in this context is the distant recoding condition that affects the global ASR performance (**?**) due to the suboptimal position of the speaker, the reverberation and the complex acoustics of the home (several rooms). However, it has been shown that this effect can be reduced with acoustic models learned on the same conditions as the target domain and using multiple channels (**?**).

Regarding the LM, voice commands for home automation systems are characterised by a small vocabulary. Although users might not agree with the grammar of the system, typical users tend to adapt their syntax to the system 'to make it work'. This behaviour is linked to the process of *alignment* where several people, when interacting with each other, tend to adapt their language to their interlocutors (**?**). This effect has been observed between human and computer as well (**?**). According to Branigan et al. (**?**), alignment occurs in HCI, but in a different way than between humans. The effect is stronger and more oriented towards enhancing communicative success (i.e., making the machine understand). It is unclear how ageing plays a role in this case. Would elderly people have more difficulties adapting to a grammar that is not their own? According to Bailey and Henry (**?**), older adults would have a reduced capacity to take the perspective of somebody else. It is unclear how this would affect the alignment process in a VUI context. Some studies have reported that older speakers tend to have more polite and longer sentences than younger speakers (**?**) while some other studies show a preference for short sentences for voice command (**?**). The need for grammar adaptation is definitely an issue that must be investigated.

The most important studies related to speech recognition in smart homes are summarised with their properties in Table **??**. From this table, it can be noticed that very few systems were implemented in realistic smart homes and tested with real users in such a context. It is clear that no automatic system was based on a voice interface and that nobody used external information for context aware control of the environment. Some studies are based on corpus analysis or Wizard of Oz use and do not involve a system running online. When end user's were included in the study, their number was always small due to the difficulty in recruiting and organising experiments with such people.

2.2 Evaluation of Voice User Interfaces for Smart Homes

Interest of elderly people regarding voice interaction at home has been studied in a few studies (**???**). In (**?**), elderly people were interested in voice command to activate windows and blinds as well as for operating television and radio. In (**?**), voice command was used for interaction during the execution of small tasks (mainly in the kitchen). The interviewees expressed their fear about a system that does not recognize what they say. According to the authors, this fear is due to the user's experience with the speech interfaces of their mobile phones. Moreover, these studies also reported that 95% of the people would continue to use the system even if it was sometimes wrong in interpreting orders. These findings are consistent with those of (**?**) in which elderly persons were questioned and put into a smart home where a voice control system was simulated by a Wizard of Oz (WoZ). They expressed high interest in voice commands for controlling the environment and the messages about security issues. They also raised concerns about the potential negative effects of such technology driving them towards a lazy life. These studies showed that the audio channel is a promising area for improvement of security, comfort and assistance in health smart homes (**?**), but it remains relatively unexplored compared to classical and mobile physical interfaces (switches, remote control, mobile phone).

Predicting user acceptability from the literature is a difficult task given the diversity of users and applications considered and the lack of consistent criteria (**?**). As can be seen in Table **??**, the criteria, evaluation methods, profile of the participants, realism of the evaluation situations are different in different studies. Thus, these studies are difficult to compare.

Study	VUI	Distant speech	Context aware	Multi-room	On-line system	Real smart home	End-user inclusion
S1	no, synthetized voice listening	no	no	no	no	no	32 aged listener
S2	browser driven by speech	no	no	no	yes	no	not operated
S3	sound (no speech)	no	no	no	no	no	corpus evaluation
S4	"Yes"/"No" dialog	micro-phone array	manual	yes	yes	no	9 young speakers
S5	WoZ dialog	no	no	no	no	no	26 aged, 25 non-aged speakers
S6	large vocabulary	no	no	no	yes	no	not operated
S7	small vocabulary	yes	no	no	no	no	corpus (5 speakers)
S8	small vocabulary	yes	no	yes	yes	realistic smart home	15 non-aged speakers
S9	self-learning VUI	no	no	no	no	virtual smart home	speakers with dysarthria
S10	speech recognition	yes	no	yes	no	realistic smart home	10 non-aged speakers
S11	dialog system	no	no	no	no	corpus	15 people with speech disorder
S12	distress detection (video and speech)	yes	no	no	no	corpus recording	4 elderly 13 non-aged
S13	VUI for smart home	yes	yes	yes	yes	realistic smart home	11 aged and visually impaired speakers

Legend:

Study	Corresponding Reference	Project
S1	(?)	MILENNIUM
S2	(?)	UBIQUITOUS COLLABORATIVE ADAPTIVE TRAINING
S3	(?)	COMPANIONABLE
S4	(?)	PERS
S5	(?)	-
S6	(?)	COMPANIONS
S7	(?)	COMPANIONABLE
S8	(?)	SWEET-HOME
S9	(?)	ALADIN
S10	(?)	DIRHA
S11	(?)	PIPIN HOMESERVICE
S12	(?)	CIRDO
S13	**This study**	SWEET-HOME

Table 1: Summary of the studies related to speech recognition in smart homes.

Evaluation	Aim	Evaluation method	Interaction method	Participant	Size
E1	Proactive home	questionnaire	RFID/speech/ motion sensor	all ages	27 people
E2	Home automation	ethnographic study	Graphical User Interface	adults	22 focus groups
E3	Home automation	interview	not used	elderly	15 people
E4	Security	questionnaire	not specified	elderly	49 people
E5	Home automation	WoZ and interview	speech/switch	elderly	200 people
E6	Emergency	Dialog system	speech "Yes"/"No"	adults	9 people
E7	Medical assistance	questionnaire	not used	elderly	82 people
E8	Home automation	WoZ and interview	speech/switch	elderly/relatives professionals	18 people
E9	AAL Context	WoZ and ASR	not specified	adults	10 people
E10	Home automation	real system and interview	speech	elderly and people with visual impairment	11 people

Legend:

Evaluation	Corresponding Reference	Project
E1	(**?**)	Proactive Computing Research Programme
E2	(**?**)	SMART HOME USABILITY and LIVING EXPERIENCE
E3	(**?**)	-
E4	(**?**)	-
E5	(**?**)	HADA
E6	(**?**)	PERS
E7	(**?**)	-
E8	(**?**)	SWEET-HOME
E9	(**?**)	DIRHA
E10	**This study**	SWEET-HOME

Table 2: Summary of the studies related to the evaluation of interactive technologies in smart homes.

- Evaluation method. Some studies used quantitative evaluations, others used qualitative ones. For instance, in (**?**), 200 Spanish people between 50 and 80 years old were questioned about different features of a smart home, but these people were not confronted with a prototype system, whereas in (**?**) the developed Personal Emergency Response System was tested with only 9 healthy young people.

- Participants. VUIs are evaluated with large samples of people or small focus groups. Furthermore, even if the systems aim at being used by elderly or impaired people, participants of studies were often healthy young adults (see Participant column of Table **??**).

- Evaluation situations and environments. Regarding the experimental settings, few experiments have actually been conducted within realistic homes and fewer within the participants' own homes with the notable exception of (**?**).

Given the complexity of the smart home domain and the importance of taking into consideration the user's personal context, *in-sitro*/*in-simu* setting (**?**) is particularly suitable to the smart home domain at the prototyping level. Of course, the frontiers between the different settings are fuzzy and actually many smart home user evaluations can be considered as having some aspects of an *in-simu* setting. Despite this diversity in experimental settings, aims, criteria, targeted users and technologies, results of these studies show convergence to some frequently expressed criteria which are described as follows.

Usability Regarding the usability, many people have expressed some apprehension towards smart home technologies because they fear not being able to use them. The system should be easy to use, easy to learn and resilient to errors (**??**). As listed by the Digital Accessibility Team (DAT)[4], smart homes can be inefficient with disabled people and the ageing population. Visually, physically and cognitively impaired people will find it very difficult to access equipment and to use switches and controls. This can also apply to the ageing population though with less severity. Thus, apart from people with hearing impairments, one of the modalities of choice is the audio channel. Indeed, audio and speech processing can give information about the different sounds in the home (e.g., object falling, washing machine spinning, door opening, foot step) but also about the sentences that were uttered (e.g., distress situations, voice commands). This is in line with the DAT recommendations for the design of smart homes which are, among other things, to provide hands-free facilities whenever possible for switches and controls and to provide speech input whenever possible rather than touch screens or keypads. Moreover, speaking is the most natural way of communication. Audio interfaces are thus *a priori* highly usable by many kinds of people in many situations (**?**).

Dependence/Confidence In some studies, people expressed concern about being dependent on such a system, especially in case of failure. Many elderly people fear that the system would break down and leave them in a critical situation by having made them dependent on the system (**??**). In (**?**), the participants emphasized that they wanted to keep control of their domestic spaces regardless of the conveniences the new technology would make available. Actually, they expressed the wish that new technologies should provide a way to be switched off so that the user always keeps control. The controllability of the smart home seems to play a crucial role for its acceptability (**?**). A smart home would thus provide much reassurance if it provided several ways of being controlled. In (**?**), among the three different interfaces provided to control a home automation system (mobile phone, media centre on TV or centralised controller on a PC) the mobile phone was the most used, but participants did use all the interfaces to control the house during the 6-month study. The study showed that the interfaces were more adapted to specific classes of actions. For instance, the mobile phone was more adapted to *instant control* (i.e., do this right now!) than *pattern control* (i.e., task automation) which was generally set using the centralised PC. The authors also pointed out that confidence in new technology is gained through the use of it, but, in general, participants were using some interfaces because they were able to check the results. Audio interfaces should thus be conceived as a complement to other ways of controlling the environment and should provide adequate feedback to the user.

Privacy/Intrusiveness Some people have expressed the wish that all these technologies do not interfere with their daily activities and that the system is as invisible as possible (**?**). In general, the participants would like to interact as little as possible with the system. It is important to note that many systems, in particular fall detectors, are relying on video cameras (**???**), but little is known about the acceptance of such sensors by the intended users who are not always included in the system design. For the elderly, there is a balance between the benefit of such monitoring (sensors of all kinds) and the intrusion into privacy. Rialle et al. (**?**) showed that the degree of acceptance of an intrusive technology varies with the severity of the pathology of the elderly person being supported. This was also confirmed in (**?**) where most of the 82 interrogated

[4]http://www.tiresias.org

participants did not think that the inquisitiveness of the system was an issue. But this study was focused only on medical applications for which, as stated before, the vital benefit of the system makes some changes acceptable in daily life. Another aspect of privacy which is emerging is what will be made of collected data. Since the system receives information of vital importance, the system has to be protected against intrusion and has to make sure that the information reaches only the right people. Smart home design must thus be respectful of privacy and should provide reassurance regarding who is going to access the collected private data (**?**).

Conclusion of this overview As summarised in Table **??**, the technologies involved in the literature are classical switches, RFID, motion sensors or graphical user interfaces to a computer. Speech technologies are sometimes taken into account but, they have not been evaluated in the context of an interactive automatic system.

3 The Sweet-Home system

To provide assistance at home through voice command, the SWEET-HOME system is composed of an audio analysis system, called PATSH, and an Intelligent Controller. The SWEET-HOME system is depicted in Figure **??**. It is linked with a home automation network composed of typical home automation data sensors and actuators (switches, lights, blinds, etc.) and multimedia control (uPnP). Moreover, the system relies on several microphones per room disseminated on the ceiling so that voice command is made possible from anywhere in the house in a hands-free way. Finally, the system contains a dedicated communication sytem that allows the user to easily contact his relatives, physician or caregiver. In SWEET-HOME, this system is *e-lio*, developed by the Technosens[5] company providing home services to elderly people through the *e-lio* box (e.g., video-conferencing, calendar, photos, etc.). In order for the user to be in full control of the system and also in order to adapt to the user's preferences, two ways of commanding the system are possible: voice command or classical tactile interfaces (i.e., switches).

Briefly, the functioning of the system is the following. All sensor data are acquired continuously by the system in an on-line manner[6]. Amongst the raw information, the raw audio data are analysed by PATSH and then speech as well as other sound events are transmitted to the subsequent stages. The Intelligent Controller continuously analyzes the streams of data and makes decisions based on these. If a vocal command is detected and, according to the context (e.g., user's location), a home automation command is generated to turn the light on, close the curtains or emit a warning message through a voice synthesizer (e.g., "be careful, the input door has remained open"). The rest of this section details the audio processing system and summarises the Intelligent Controller which performs the context aware decision.

3.1 The audio processing system: PATSH

The audio processing system PATSH is depicted in Figure **??**, it is composed of several processing modules organised into a pipeline, the modules perform the following tasks: (1) multichannel

[5] `http://www.technosens.fr/`

[6] Here, on-line means that each time a new event appears, it is immediately queued for processing.

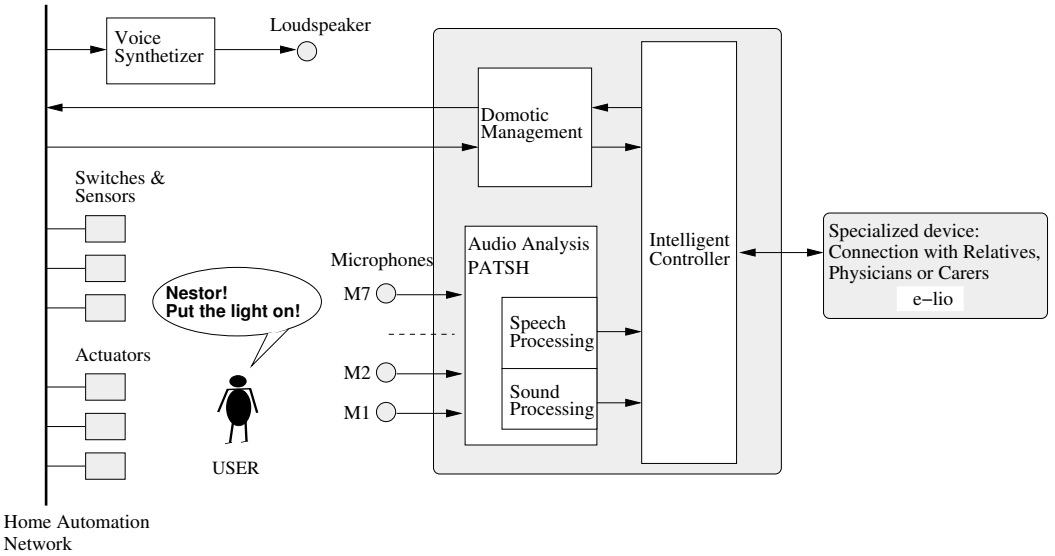


Figure 1: The SWEET-HOME system: example of a user asking to switch the light on.

data acquisition (16bits, 16kHz, 7channels), (2) **sound detection**, (3) **sound/speech discrimination**, (4) **sound classification**, (5) **Automatic Speech Recognition** (ASR) and extraction of voice commands, and (6) **presentation**, communicating the results to the Intelligent Controller. The modules exchange data through instances of a **sound object**. Each sound object contains a segment of the multidimensional audio signal whose interpretation is continuously refined along the processing pipeline. PATSH deals with the distribution of the data among the several modules that perform the processing to interpret the audio events. For a complete description of the system, the reader is referred to (**?**), only the most important modules are briefly presented here.

The multichannel **data acquisition** and **sound detection** modules are tightly coupled. The input channels are acquired synchronously and each time a buffer is completed, a sound object containing the samples of all the channels is sent to the detection module. The sound detection module processes each channel independently in parallel (i.e., multiple instances of the sound detection module), keeping a local buffer of the signal being processed. Each time an event is detected by one of the sound detection instances, a new sound event is created and sent to the subsequent stages. In case a sound event is detected simultaneously on different channels, the **Sound object** with the best Signal to Noise Ratio (SNR) is kept. The detection of the occurrence of an audio event is based on the change of energy level of the 3 highest frequency coefficients of the Discrete Wavelet Transform (DWT) in a sliding window frame (2048 samples without overlapping). Each time the energy on a channel goes beyond a self-adaptive threshold, an audio event is detected until the energy decreases below this level (**?**). At the end of the detection, the SNR is computed by dividing the energy in the event interval and the energy in previous windows outside this interval.

Once sound occurrences are detected, the most important task is to distinguish speech from other sounds. Therefore, the **sound/speech discrimination** module has a crucial role: firstly, vocal

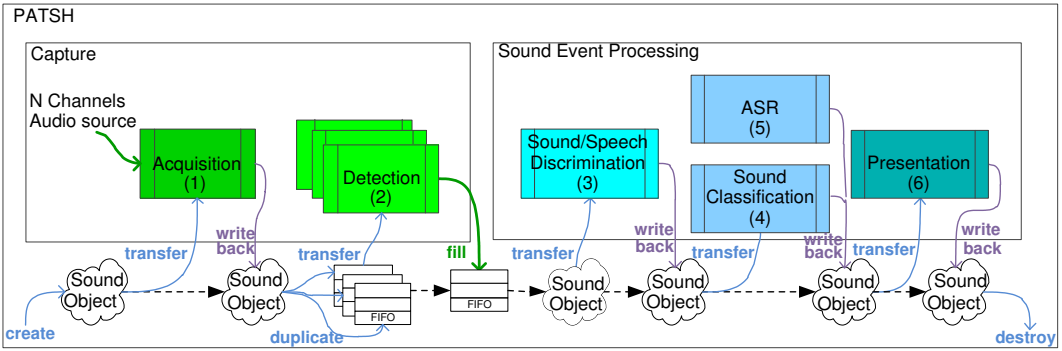


Figure 2: The PATSH architecture [Vacher et al., 2013]. On the top, the different modules of the process pipeline. At the bottom the sound object exchange pipeline.

orders must not be missed, secondly, non-speech sounds must not be sent to the ASR because undesirable sentences could be recognized. The method used for speech/sound discrimination was to train a GMM (Gaussian Mixture Model) for each class (sound or speech) for which each input sound event is described by a vector of 16 MFCC (Mel Frequency Cepstral Coefficients) feature vectors extracted from 16ms signal frames with an overlap of 8ms. The GMMs were trained separately for each class on a dataset composed of examples of typical sounds in a home and typical sentences (see (**?**) for details about the dataset). The same technique was used for the **sound classification** module but with a larger number of classes. In addition, to recognize only vocal orders and not every sentence uttered in the home, all sound events shorter than 150 ms and longer than 2.2 seconds were discarded as well as those whose SNR is below 0 dB. These values were chosen after a statistical study on a dataset (**?**). Once the sound event is classified as speech (resp. sound), the sound object is sent to the **ASR** module to extract voice commands (resp. sound classification module). Due to the focus of the study on the speech interaction, the sound classification will not be detailed in this paper (see (**?**) for details about the sound classification) while the ASR module is detailed in chapter **??**.

Once the ASR or the sound classification is performed, the **presentation** module translates the sound objects into an XML representation containing: the type of sound (every day life sound or speech), the class of sound or transcript of the speech, the SNR, the duration and the time of occurrence as well as the microphone source. This XML representation is sent to the intelligent controller through a SOAP connection.

3.2 Speech recognition and voice command recognition

For a voice command application the ASR module must quickly decode the speech events so that transcriptions are sent as soon as possible to the intelligent controller. That is why the Speeral tool-kit (**?**) by the LIA (Laboratoire d'Informatique d'Avignon) was used. Indeed, its 1xRT configuration allows a decoding time similar to the signal duration. Speeral relies on an A^* decoder with HMM-based context-dependent acoustic models and trigram language models. HMMs are classical three-state left-right models and state tying is achieved by using decision trees. Acoustic vectors are composed of 12 PLP (Perceptual Linear Predictive) coefficients, the

energy, and the first and second order derivatives of these 13 parameters.

The acoustic models of the ASR system were trained on about 80 hours of annotated speech. Furthermore, acoustic models were adapted to the speech of 23 young speakers recorded in the same home during previous experiments by using Maximum Likelihood Linear Regression (MLLR) (**?**), in order to fit to the acoustic condition of the home. In a final step, this acoustic model was adapted to each participant's voice thanks to a short text read by the participant before the beginning of the experiment.

A 3-gram Language Model (LM) with a 10k lexicon was used. It results from the interpolation of a *generic* LM (weight 10%) and a *domain* LM (weight 90%). The *generic* LM was estimated on about 1000M of words from the French newspapers *Le Monde* and *Gigaword*. The *domain* LM was trained on the sentences generated using the grammar of the application (see Fig. **??**). The LM combination biases the decoding towards the *domain* LM but still allows decoding of out-of-domain sentences. A probabilistic model was preferred over strictly using the grammar because it makes it possible to use uncertain hypotheses in a fusion process for more robustness.

Possible voice orders were defined using a very simple grammar as shown on Figure **??**. Every command starts with a unique keyword that indicates whether the person is talking to the smart home or not. In the following, we will use *'Nestor'* as the keyword. The grammar was built after a user study that showed that targeted users prefer precise short sentences over more natural long sentences (**?**). In this study, although most of the seniors spontaneously controlled the home by uttering sentences, the majority said they wanted to control the home using keywords. They believe that this mode of interaction would be the quickest and the most efficient. This study also showed they had a tendency to prefer or to accept the 'tu' form (informal in French) to communicate with the system.

The last step of the ASR was composed of a voice command recognizer. Briefly, the best of the ASR output hypotheses was phonetized and a distance was computed against all the possible phonetized sentences of the grammar. For each comparison, the minimal Levenstein distance was computed using Dynamic Time Warping (DTW). If the distance was above a certain threshold then a voice command was detected otherwise it was rejected. This approach permits recovery from some decoding errors such as word declination or light variations (the blind, the blinds, etc.). In a lot of cases, a miss-decoded word is orthographically close to the good one (due to the close pronunciation).

```
basicCmd         = key initiateCommand object |
                   key emergencyCommand
key              = "Nestor"
initiateCommand  = "ouvre" | "ferme" | "baisse" | "éteins" | "monte" | "allume" | "descend" |
                   "appelle" " | "donne"
emergencyCommand = "au secours" | "à l'aide"
object           = [determiner] ( device | person | organisation)
determiner       = "mon" | "ma" | "l'" | "le" | "la" | "les" | "un" | "des" | "du"
device           =  "lumière" | "store" | "rideau" | "télé" | "télévision" |
                    "radio" | "heure" | "température"
person           = "fille" | "fils" | "femme" | "mari" | "infirmière" | "médecin" | "docteur"
organisation     = "samu" | "secours" | "pompiers" | "supérette" | "supermarché"
```

Figure 3: Excerpt of the grammar of the voice orders (terminal symbols are in French).

3.3 The Intelligent Controller for context aware interaction

The reasoning capabilities of the system were implemented in the Intelligent Controller depicted in Figure **??**. The Intelligent Controller is represented by the upper box. It gathers streams of data from the external systems and transmits orders back to the home automation network. All these streams of information are captured and interpreted to recognize situations and to determine the context before making decisions. More precisely, these input data are composed of asynchronous events (e.g., infra-red sensors, motion detector, door opening states of some devices), time series (e.g., temperature, water consumption) and recognized voice commands sent from PATSH.

When the user asks to turn on the light "Nestor turn on the light", he does not specify the desired level of illumination (i.e., strong or soft) or which lamp to be lit (e.g., bedside or ceiling lamp). For the user, this information is implicit and indeed another human being would probably infer correctly the user's goal. But for the system, this lack of information must be recovered from its knowledge of the context: the controller must infer the participant's location to switch the light on in the appropriate room and must infer her activity to determine the appropriate illumination and lamp in the room (e.g., if she is asleep then the bedside lamp might be more appropriate than the ceiling one). In this example, the context is composed of two inferred parameters from sensor data: the location and the activity.

The estimation of the current context is carried out through the collaboration of several processors, each one being specialized in a specific source of information. All processors share the knowledge specified in ontologies and use the same repository of facts. Furthermore, the access to the knowledge base is executed under a service oriented approach that allows for any processor to be notified only about particular events and to make inferred information available to other processors. This data and knowledge centred approach ensures that all the processors use the same data structure and that the meaning of each piece of information is clearly defined among all of them.

The main aspects to be considered for context aware decision making are the location of the inhabitant and the current activity. These kinds of information are useful to eliminate ambiguity in the decision making process. Other works have also reckoned location and activity as fundamental for context-aware inference (**??**). In order to perform location and activity inference, two independent modules were developed and integrated into the framework. The first applies a two-level dynamic network method able to model the links between sensor events and location assumptions. Data fusion is achieved by spreading activation on the dynamic network. The second module uses a classifier, based on Markov Logic Networks (MLN), to carry out activity recognition. Due to space limitations the reader is referred to (**?**) for further details.

In SWEET-HOME, the actions that the intelligent controller can make are the following:

- turn on/off the {light, radio}
- close/open the {blinds, curtains}
- give the {temperature, time}
- warn about {open windows, unlocked door}
- order the *e-lio* system to call a specific number or to send out an emergency call.

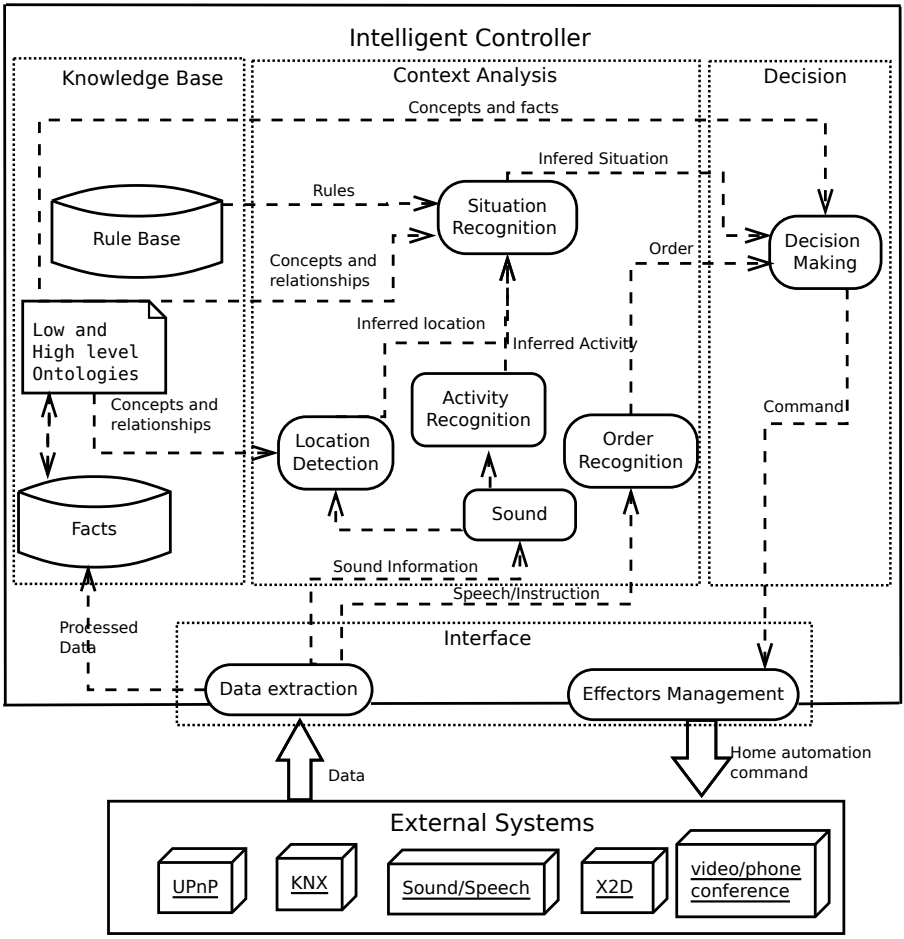


Figure 4: The Intelligent Controller Diagram.

These actions constitute a subset of a larger set of possible actions resulting from a previous user study (**?**). This study showed that the users (in this case, elderly people) were more interested in actions providing security and avoiding dangerous manipulations. Actions saving time or related to food were not liked by the participants. For instance, automation of coffee machine was unanimously disliked since seniors want to keep coffee making as they have time to do so and it is part of a social activity (making coffee for their visitors). Of course, this set of actions must be adapted to every user and home, but this predefined list was useful for the evaluation of the system.

In order to make the "best" decision from uncertain data, the decision reasoning was implemented using an influence diagram approach. Influence diagrams (**?**) are probabilistic models used to represent decision problems. They extend Bayesian networks – composed only of state nodes – by the inclusion of two other types of node: action and utility. An action node is a variable corresponding to a decision choice (e.g., turning the light on or warning the user). The state nodes represent the variables of the domain that are affected by the actions. Finally, utility nodes

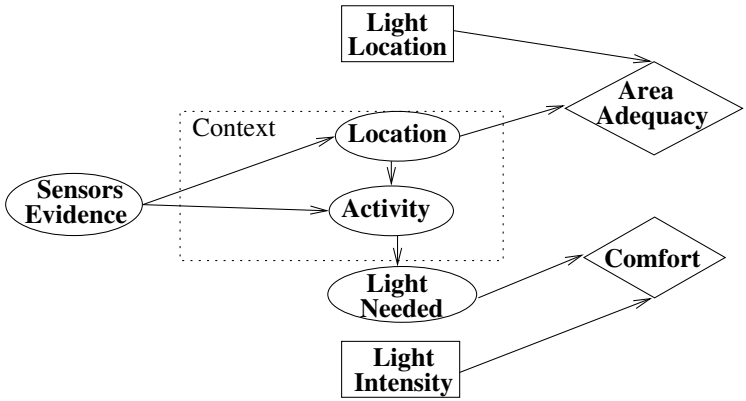


Figure 5: Influence diagram for a decision after a vocal order is recognised.

are variables that represent the utility value obtained as a consequence of applying the decided actions. For instance, turning the light on at full intensity when the person is asleep would have a negative utility.

Making a decision with an influence diagram typically consists of computing the Expected Utility $EU(a) = \sum_x P(x|a,e)U(x)$ for every action $a \in A$ from a set of state variables x which are influenced by the action a, the set of input evidence e and $U(x) \in \mathbb{R}$ the utility of x. The chosen action is the one bringing the highest expected utility.

Figure **??** shows an example of an influence diagram, where a decision is made as a response to a vocal order *Turn on the light*. In this case, the setting of action variables, represented by rectangular nodes, indicates which *lights* are operated and their *intensity*. Oval nodes are the state nodes, some of which are affected by the decision, while the others belong to the context (within the dashed area). Two utility nodes influence directly the decision: the *comfort* of the inhabitant and the suitability of the activated *light location* that ideally should be the same as the inhabitant's one. Note that this location is not easy to determine in some cases since the inhabitant could be moving in the home while uttering the vocal order.

In SWEET-HOME, the influence diagram was implemented using a statistic-relational approach so that decision rules can be expressed in first order logic while the reasoning takes incompleteness and uncertainty into account. For instance, let's consider the following rules in the MLN formalism:

...
3.35 $LightLocation(l1) \wedge Location(l1) \rightarrow RightArea(good)$
0.12 $LightLocation(l1) \wedge Location(l2) \wedge NextTo(l1,l2) \rightarrow RightArea(acceptable)$
...

The first rule expresses that if the chosen action is to turn on the lamp located in $l1$ and that the user is also in location $l1$ then the state variable $RightArea$ takes the value $good$. The number 3.35 translates the fact that the rule is very often true. The second rule states that if the chosen lamp is close to the location of the inhabitant then it is an acceptable choice. However, its

weight is low 0.12 meaning that the rule is rarely true. In MLN the influence diagram is a set of weighted formulae. The weights are learned from data and are relative to each other. In the case of the example, the first rule is true for a majority of the time in the training dataset while the second rule is true only for a few cases. A thorough description of the decision process is far beyond the scope of this paper, but the above description of the influence diagram and its implementation in the MLN formalism provides the reader with the gist of the overall approach in order to understand the paper (e.g., how the location is used to make a decision). For more details about the intelligent controller, the reader is referred to (**?**).

Decision models are available for each kind of action that are related to an object that may exist at different places. For instance, when a voice command such as 'turn on the light' is received, then the decision about the lighting is executed, when it is about the blinds the decision model about the blinds is run and so on.

4 Protocol

To assess the system with targeted users, a user study was set up consisting of semi-directed interviews and sessions in which each participant, alone in a flat, interacted with the system following predefined scenarios. This section describes the objectives of the assessment, the profile of the targeted participants, the experimental setting as well as the procedure that was applied with every participant.

4.1 Objective of the assessment

The experiment was conducted at the end of the SWEET-HOME project and had two major assessment objectives: usage and technique.

The main objective was to evaluate the system regarding 1) the user's *interest* in the solution being evaluated, and 2) the *accessibility*, *usefulness* and the *usability* of the system. This evaluation was also a way to identify the main impediments to the appropriation of the system and ideas of improvement that could emerge from the user study either directly suggested or through the collected evidence.

To avoid a user evaluation biased by technical performance issues, a technical validation of the system was performed four months before this experiment (**?**). It showed the adequacy of the system to the task and some of its limitations. For instance, the voice command error rate was acceptable but the audio processing had difficulties in handling a large number of sound events (e.g., a western cavalry charge on TV) and had a very slow response time (up to four seconds). These problems were addressed before this user evaluation (for more details about the previous technical evaluation the reader is referred to (**?**)). However, due to time constraints the previous technical evaluation was assessed only with typical users (i.e., mostly 'naive' colleagues from the lab). Though the main technical problems have been addressed, the technical performance of the system with the targeted users was also assessed in order to compare objective measures (e.g., word error rate, time response) with subjective ones (e.g., preferences, opinions).

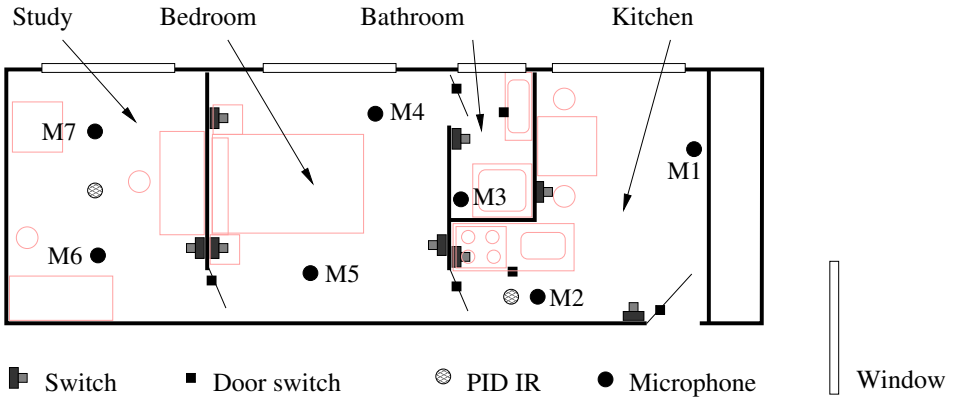


Figure 6: Position of the microphones and other sensors inside the DOMUS smart home.

4.2 Participant profile

The targeted users were seniors and people with visual impairment.

The senior profile was any person living alone in an independent non-hospitalised accommodation. The minimum age was set to 75 years old[7] but was then reduced to 74 due to the difficulty of recruiting participants. The focus of the study was to target seniors who were on the edge of loosing some autonomy not seniors who had already lost their autonomy. In other words, we sought seniors who were still able to make a choice regarding how the technology could help them in the near future in case of any degradation of their autonomy.

The visually impaired category was composed of adult people living either alone or in couple and whose handicap was acquired after their childhood. No upper age limit was given. According to the French regulation a person with low vision is someone whose visual acuity of the best eye after correction is lower that 4/10. They should not be confused with blind people who can not see at all. The targeted users in our case could still see but with very low acuity.

4.3 Experimental settings

The whole experiment was run on the DOMUS platform of the *Laboratoire d'Informatique de Grenoble*. The platform is composed of a set of rooms equipped for studying, conceiving and assessing smart technologies. Figure **??** shows the details of the smart home of the platform. It is a thirty square meters suite flat including a bathroom, a kitchen, a bedroom and a study, all equipped with 150 sensors and actuators. The flat has been equipped with 7 radio microphones set in the ceiling and with the SWEET-HOME system. A PC was running the PATSH system and another the intelligent controller. *e-lio*, the communication device used to initiate a communication between a senior and his relative, was placed in the study.

[7]This threshold is the same as the one used by the INSEE `http://www.insee.fr/fr/themes/tableau.asp?reg_id=0&ref_id=NATnon02150`

4.3.1 Scenarios

To validate the system in realistic conditions, 4 scenarios were designed in which every participant was placed in the following situations:

1. “You are finishing your breakfast and you are preparing to go out”
2. “You are back from shopping, you unpack the shopping and put it away, you would like to have some rest”
3. “You want to communicate with your relatives”
4. “You are waiting for friends who are about to come”

Each of these scenarios is designed to last between 5 to 10 minutes but there was no constraint on the execution time. Scenario 1 and 2 were designed to make the user perform daily activities while uttering voice commands. Figure **??** shows the details of the first scenario. The participant was provided with a list of actions to perform and voice commands to utter. As it can be seen, the voice commands were provided but not the grammar. Indeed, the grammar would have been too difficult to manipulate for people unfamiliar with this format.

Go to the kitchen
Ask for the ambient temperature:
Nestor donne la température (*Nestor give the temperature*)
You can have a snack
Once finished, put the dishes in the sink
Ask for the current time:
Nestor donne moi l'heure (*Nestor give the time*)
You realise that it is late, you must go shopping
before leaving the home, you want to turn off the light:
Nestor éteins la lumière (*Nestor shut off the light*)
you also want to close the blinds:
Nestor baisse les stores (*Nestor close the blinds*)
Finally you leave the home

Figure 7: Excerpt of the first scenario given to the participant.

Scenario 3 was devoted to the command of the *e-lio* system and to the simulation of emergency calls. Scenario 4 contained much more freedom. Participants were told to generate a random number of voice commands without predefined sentences. This scenario was placed at the end of the others to test whether the participants would naturally adhere to the grammar or not. These scenarios allowed us to process realistic and representative audio events in conditions which are directly linked to usual daily living activities.

Each participant had to use vocal orders to switch the light on or off, open or close blinds, ask about the temperature and ask to call his relative. The participants were told to repeat the order up to 3 times in case of failure. A wizard of Oz was used in case of persistent problems.

In the experiment, the participants did not strictly follow the scenarios and in particular did not perform all the possible daily activities. Indeed, some activities such as lying on the bed or

washing the dishes were not well accepted by the participants, that is why the daily activities in this experiment were very restricted, and hence their analysis are not included in the study.

4.3.2 Interviews with participants

Two interview phases were planned. A first interview at the beginning of the session and a debriefing at the end. Both interviews were semi-guided consisting of predefined questions and possible answers but the participants were free to express any opinion or details they wanted. The first interview consisted of 10 questions related to the habits, challenges and the opinion of every participant about his own home. This first exchange helped to make him more confident as well as to understand his familiarity with ICT. At the end of the session, the ergonomist debriefed with the participant on his feeling about the experiment. The aim was to understand his behaviour and to assess the acceptability of the system. Four topics related to the scenarios were identified. The first was about the overall assessment of the system (interest, interaction mode, difficulties), the second was related to the video-conferencing system, the third focused on the grammar used to communicate with the system and the last explored the expectations and perspectives of the system through other features. 30 questions were asked in total.

4.4 Schedule of one session

Each participant was invited to come with one of his relatives or friends. At his arrival, he was welcomed by one of the experimenters who explained the aim of the experiment, what will be done with its data and how he can access that data. Then he was asked to give a signed informed consent. The participant visited the smart home to ensure that he would find all the items necessary to perform the scenarios. It was necessary to explain the right way to utter vocal orders and to use the *e-lio* system. At the end of the visit, the participant was asked to read a text of 25 short sentences in order to adapt the acoustic models of the ASR. During the adaptation of the acoustic models, the participant and his relative had the first interview.

Once the participant, the acoustic models and the experimenters were ready, the participant played each scenario with a short break between each to discuss with the experimenter. Each participant was given a sheet of paper with the scenario in which the voice commands were emphasized. In the case of users with visual impairment, either the scenario was written using very big fonts or the instructions were given using a mobile phone (only one person). For scenario 3, the relative was asked to answer the video-conferencing system when the participant called. In case there was no relative, an experimenter took this role (this was agreed during the visit).

Finally the session ended with a second interview concerning the feeling of the user about the system.

Overall a session was planned to last about one hour and 30 minutes. Sessions were not organised specifically for one group but according to the availability of every participant.

5 Data

The experiment was run between April and May 2013 and involved 11 participants. This section describes the participants that were recruited and the recorded data.

5.1 Participants

The participants were 11 people from the Grenoble area. They were divided up into two groups: seniors (n=6); and people with visual impairment (n=5). Table **??** and **??** summarise the participants' characteristics.

Table 3: Summary of the senior group.

Participant	gender	age	habitation	health problem	phone	computer
S1	female	91	residential home	weak sight, rheumatism	landline	no
S4	female	82	flat	high blood pressure	landline& mobile	laptop
S6	female	83	residential home	weak heart	landline& mobile	no
S7	female	74	flat	stroke in 2004	landline& mobile	PC
S9	female	77	flat	weak heart, blood pressure problem	landline& mobile	laptop
S11	female	80	flat		landline& mobile	no

The mean age of the elderly group was 81.2 years old (SD=5.8), and all were women. These people were single and lived in a flat in full autonomy. Three of them lived with pets. Four of them had some health problems due to ageing including S7 who had a stroke, reducing her mobility. Some of the participants revealed that they had adjusted their house to their condition (age, disability, finance) going from the installation of an emergency call system to the point of moving house. These adjustments seem to address the fear of falling and a sense of insecurity. The homes of the seniors included standard white goods, landline and often mobile phone. Some of them had a computer for e-mail and browsing the web. The mobile phone was only used for communicating outside the home (no SMS, no browsing over the web).

Table 4: Summary of the visually impaired group.

Participant	gender	age	habitation	status	adjustement
S2	female	67	flat	couple	speech synthesizer
S3	male	50	flat	single	home automation system, house arrangement
S5	male	66	house	single	pathway illuminated by LED
S8	female	64	flat	couple	purchase of a ground floor flat
S10	male	64	flat	couple	no information

The mean age of the visually impaired group was 62.2 (SD=6.9), and 2 out of 5 were women. None of them lived with children at home. The impairment was not congenital but was caused by accident or disease. The progress of the impairment was unknown. The participants made some adjustments to their home which were mostly related to technological evolutions (illuminated pathway, home automation system). Moreover, they said they have difficulties when an object is not at its usual place, thus, all visitors must replace every object to its right place. The home adjustment was mostly an "augmentation" of the devices, they did not change the house but adapted its functioning. For instance, one participant reported "I have a normal phone, the only difference is that it speaks. It is the one of the man on the street".

All participants (elderly and visually impaired) had regular visits and remained involved in society

Table 5: Recorded audio data.

Speaker ID	Group	Age	Sex	Duration	Nb. of speech utterances	Nb. of commands	SNR mean (dB)
S01	Senior	91	F	24mn 00s	59	37	16
S02	Visually impaired	66	F	17mn 49s	67	26	14
S03	Visually impaired	49	M	21mn 55s	53	26	20
S04	Senior	82	F	29mn 46s	74	27	13
S05	Visually impaired	66	M	30mn 37s	47	25	19
S06	Senior	83	F	22mn 41s	65	31	25
S07	Senior	74	F	35mn 39s	55	25	14
S08	Visually impaired	64	F	18mn 20s	35	22	21
S09	Senior	77	F	23mn 05s	46	23	17
S10	Visually impaired	64	M	24mn 48s	49	23	18
S11	Senior	80	F	30mn 19s	79	26	23
All	-	-	-	4h 38mn 59s	629	291	-

by participating in social activities. However, participants with sight impairment often stayed at home, an environment they knew and mastered well. Finally, from the data collected, it can be assumed that none of the participants were definitely technophobic.

5.2 Technical data

Before the beginning of the experiment, each participant signed a consent form enabling researchers to use the data for research purpose only. For each participant, any problem encountered during the experiment was recorded in the notebook of experimentation. All the data streams were processed online while a participant was interacting with the smart home. The recorded data available for each participant were:

- audio recording of a read text for adaptation purpose: 1 file;
- the acoustic model resulting from the adaptation to the participant with this text; this model was used by the ASR system during the experiment with the corresponding participant;
- questionnaire for user study: 1 file;
- video traces: 6 files used for annotation of the activity and of the localisation of the participant in the home;
- home automation traces (home automation devices, Intelligent Controller and wizard of Oz): 1 file;
- 7-channel raw audio signals: 7 files;
- a directory containing the data corresponding to the sound objects analysed by PATSH: for each sound object, the audio file extracted online and its associated XML file with analysis results (beginning and end of the audio signal, Signal to Noise Ratio, date and time, signal duration, discrimination result: speech or sound, recognition result: sentence hypothesis or sound class).

The recorded audio data are summarized for each participant in Table **??**. Regarding S11, PATSH crashed 17mn and 25s after the beginning of the experiment and had to be restarted; therefore, approximately 1 minute of audio recording was lost. These data are part of the `Sweet-Home Corpus` (**?**).

5.3 Annotation of the data

During the experiment, the seven audio channels were continuously recorded. The transcription and the annotation of the audio signals were made on two channels (kitchen and study) that were gain adapted and mixed. The microphones corresponding to these two channels were optimally placed to capture the speech utterances of the experiment. The speech transcription was performed using transcriber (**?**) a famous tool in the speech processing community.

Once the speech transcription was performed, each utterance was semantically annotated using semantic frames (**?**) of the following type:

Frame	Slot	Value
open_close	object	{blinds, curtains}
	action	{open, close}
activation	object	{light, radio}
	action	{on, off}
call	recipient	{daughter, husband, ...}
question	object	{temperature, time}
distress		

It must be emphasized that the labelling was performed for every utterance for which the participant's "intention" was to utter an order. However, the participant did not always follow the grammar and sometimes repeated several times the same order. Thus, in addition to the semantic labelling, each home automation order was also annotated with respect to the `syntax` $\in \{yes, no\}$ and the rank of the `repetition` $\in \{0, 1, 2, \ldots\}$.

Furthermore, when the utterance was not a home automation order, it was categorised into one of the following classes {spontaneous speech, noise, synthetic voice}. This annotation was entirely performed by one author and checked by another author. In case of disagreement, a consensus was reached.

The location was marked up using the video and the Advene software using the same procedure as in (**?**).

6 Analysis

This section presents the analysis of the results of the experiment both in terms of objective measurements (Sections **??**, **??**, **??** and **??**) and subjective feedback (Sections **??** and **??**).

6.1 Performance of the system

This section reports the performances of the system in term of context recognition, speech recognition, voice command recognition and decision.

The most important element of the context is the localisation (i.e., the ability for the system to infer which room the person is in). The localisation performance was 85.36% (SD=3.85%) accuracy on average per subject computed every second in each record. This means that nearly 15% of the localisation was wrong. However, when the localisation performance is computed only at the time at which a command is generated following a recognised voice command (i.e., the location used by the decision module) the performance reaches 100%. This means that the system was always making a decision using the correct location information. Since the participants did not act daily activities (cf. section **??**), the activity recognition was not assessed in this study.

Regarding the speech events, during the experiment, there were 629 utterances. 211 were home automation orders (34%), 40 were questions about temperature or time (6%), 40 were distress calls (6%), 66 (10%) were actually generated by the speech synthesizer, 10 (2%) were noise occurrences wrongly classified as speech and 262 were other spontaneous speech occurences (42%, mostly during the video-conferencing with a relative). Only 29 speech utterances were missed (4%), but 85 of the detected ones were rejected (14%) either because they were below the minimum SNR threshold or because they were out of the acceptable duration range. Therefore, 18% of the utterances were not treated by the system.

The word error rate (WER) of the voice commands that were syntactically correct was computed and led to 43.23%. Under standard conditions, this WER is not a very good result but considering the constraints of the application (distant speech) and the nature of the users (atypical speech) this score is not considered as a failure. As a matter of comparison, the speech utterances were fed to a well known publicly available speaker independent ASR API available on-line and the resulting WER was 69.1%. It is also worth noticing that other spontaneous speech events were not well decoded by the system (91% WER in the original PATSH ASR) while the ASR API available on-line gave better results (64% WER). Though this comparison must be taken with caution since the ASR API was not designed for the task, this emphasizes the difficulty of the task. When the WER is computed on non-voice command speech the performance dramatically decreases. This is partly due to spontaneous speech, the distant condition and the language model which was not adapted. This is a very good behaviour of the ASR since only the voice commands must be captured by the system and not the personal conversations. The ideal situation would be a 0% WER for voice commands and a high WER with other kinds of speech.

When considering only the correctly uttered voice commands, 41% of them were badly recognized and then not detected. There was no confusion among the detected orders (e.g., switching the light on while the order concerned the blinds) leading to a performance close to 100% precision. We believe it is better to favour the precision over the recall as it is safer to ask the user to repeat than having a home misbehaving.

Overall, there was a ratio of voice command repetitions of 76.2% with 93.6% for the seniors (i.e., almost all voice commands were repeated once) and 55.4% for people with visual impairment. Figure **??** shows the ratio of missed voice commands per participant ordered by age. A miss was considered when the participant uttered a voice command respecting the grammar but when the corresponding action was not activated by the system. Thus, any uttered sentence not respecting the grammar was not regarded as an error of the system. It can be seen that the

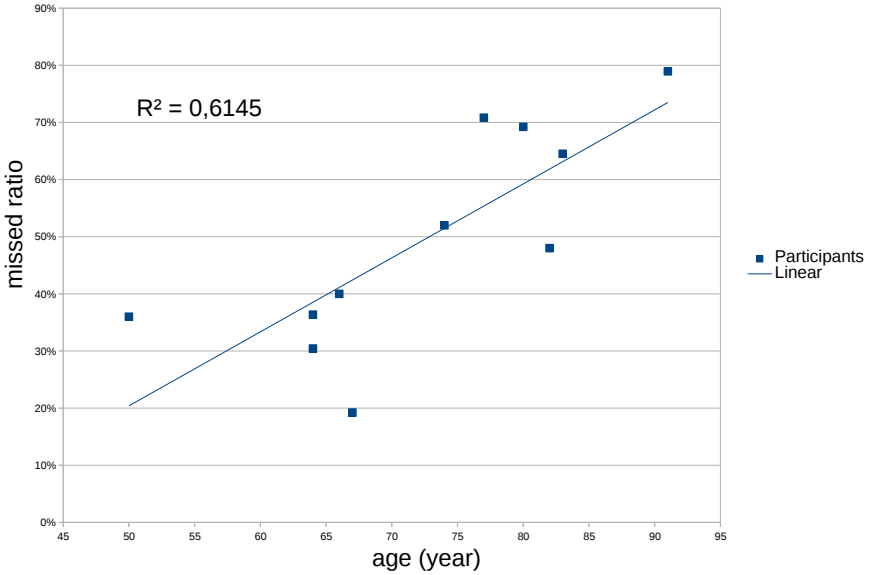


Figure 8: Missed voice command ratio as a function of the age of the participants.

older the person, the higher the error rate. It can also be observed that at 45%, there is a perfect separation between the visually impaired group and the senior group (all participants with visual impairment were below 70 years old). Though this is a small data set, this trend is perceptible in most of the following analysis.

Regarding the decision phase, among all the speech events that were sent to the intelligent controller, 141 were correctly recognized commands, while only 2 were confusions (between "give the time" and "give the temperature") and 3 others were false alarms. For 2 of the false alarms this was due to bad decoding provoked by hesitations in the utterance. These hesitations inserted phonemes in the utterance making the sentence closer to the phonetised voice command grammar. These two utterances were interpreted as "Nestor give the time". The last false alarm appeared after a period of music and once again, the correct time was given through the TTS.

6.2 Analysis of the user's behaviour in front of the system

Videos of the participants were analysed in order to observe behaviour patterns. Two recurrent patterns of sequential uttering of commands were observed. Participants with visual impairment strictly followed the sequence of commands to achieve. They uttered a command only when the previous one was completed. Moreover, their attention was totally focused on the voice command task and they never undertook any other activity at the same time. By contrast, seniors uttered orders even if the action resulting from the previous one was not completed (e.g., uttering the next command while the blinds were still making noise). At a lower level, different patterns were observed during two different steps of the vocal command: when the participant was saying the

order and while the system was processing the command. These are detailed below. Then, the relationship between the participants and the system during the experiment time is analysed.

Participants' behaviour while uttering an order Once they understand what the order to be uttered is, participants with visual impairment paid attention to the object they intended to control (such as a lamp in the case of "turn off the light" order). This attention was manifested by a move (S8, S10) and/or a change of gaze direction (for S2, S3, S5, S10) towards the object. Senior participants did not follow the same behaviour. They uttered the orders as soon as they became acquainted with them, and most of them did not move or look at another thing other than the written orders (S4, S11, S6).

Participants' behaviour after the voice command utterance Several commands had a long time execution. For example, closing the blinds took 23 seconds. During this time, participants with visual impairment just waited and did not do anything else (S5, S8, S10). For example, when S8 ordered the blinds to be closed, he was sitting on a chair with the window behind him. Once he uttered the voice command, he waited without moving till the end of the operation. Seniors did not systematically wait for the command completion to move. They continued their activities during the command execution. Systematically, once the command was completed, they checked it.

Relationship between the participants and the system The system listened to participants' commands through microphones hidden in the ceiling. The participants were left alone in the flat during all the scenarios. It was observed that this configuration (to speak into the void) was awkward for seniors. They looked for an interlocutor when they uttered commands as well as when the smart home answered a question. For instance, when the smart home gave the temperature through TTS, S7 and S9 moved to the room where the voice was generated. By contrast, participants with visual impairment did not look for an interlocutor and mostly only paid attention to the execution of their commands.

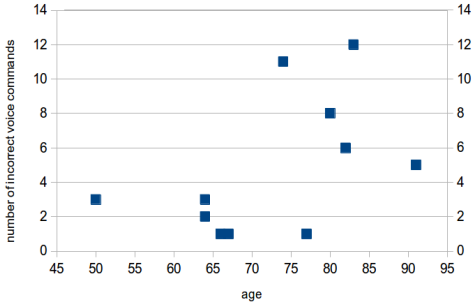
6.3 Adaptation of the participants to the voice command grammar

The voice commands had to strictly follow a fixed grammar (cf. Section **??**). For example, the first word of a voice command had to be "Nestor". However, some of the participants complained about the grammar and deviated from it. It is interesting to analyse this deviation to set up future forms of user grammar adaptation. This analysis has been performed on the last scenario of the experiment. In this scenario, participants were asked to control the home without an explicit description of the voice command. Despite the 3 previous scenarios where participants had to respect and learn the grammar, this fourth one shows some personal adaptation of the voice command.

The transcription of the sentences uttered when the participant's intention was to command the home was studied by a textual analysis. Of the 90 uttered words, 21 belonged to the grammar. As shown in Figure **??**, participants used different variations of the same verbs to express voice commands. The infinitive form (such as "éteindre" *to turn off* instead of "éteins" *turn off*) was systematically used by S4, S9 and S10. For instance "Nestor, to switch off the radio" instead of

Term	Frequency	Term	Frequency
la	40	s'il	5
Nestor	37	plait	5
les	22	monter	5
radio	21	allumer	5
stores	19	si	4
lumière	19	pas	4
éteins	16	musique	4
monte	11	moi	4
allume	11	ah	4
vous	8	tu	3
éteindre	7	pouvez	3

(a) Frequency of uttered words.



(b) Voice command syntax error number vs age of the participants.

Figure 9: Frequency of the terms and number of syntax errors.

"Nestor, switch off the radio". This phrasing is quite unnatural but is culturally associated with the phrasing of a talking machine. A senior (S11) spoke to the system by using courtesy form. Thus, she used the courtesy form of verbs such as "Pouvez vous éteindre la lumière ?" *Could you turn the light off?* She did it from the first command and she followed the same scheme (infinitive or courtesy form) till the end of the scenario. This shows that the verbal form employed by the participants is not linked to the task but seems to reflect how the user considers the system.

In addition to the courtesy verbal form observed with S11, other courtesy marks were observed. Three participants (S4, S10 and S11) added, at least one time, "s'il vous/te plait" *please* to their orders. For these participants, the system is not only a tool but an interlocutor to interact with.

This feeling has been encouraged by the fact that orders begin with the keyword "Nestor"[8] which is a well known French given name and all participants identified it as a human name (the name Nestor is rarely used with pets). In interviews, participants were asked what they thought about the use of "Nestor". No senior questioned the use of a name to address the system but half of them (S6, S9 and S11) as well as S3 would like to customize it. Two persons with visual impairment (S5 and S10) would prefer no name to communicate with the system (S10 used the infinitive form of verbs in voice command). When the system name is disliked, the opinion of people with visual impairment is stronger than the opinion of seniors. Seniors want to change the name because they dislike its sound or because they prefer another name, whereas people with visual impairment want to suppress the name or change it because it is not adapted to their daily usage. During the fourth scenario, no participant used another name than "Nestor" but S4, S6, S9, S10 and S11 frequently did not mention it in the voice command. Thus, though only people with visual impairment complained about the keyword, 4 of the 6 seniors (S4, S6, S9 and S11) did not use it when they were not guided.

Figure **??** shows the number of syntactic errors made by the participants when uttering an order over the whole experiment. Only one senior (S1) followed the command sentence structure. One of the main causes of not following the grammar was the absence of the keyword "Nestor". Indeed seniors often spoke to the system as if they were discussing and thus did not feel the need to specify a name at the beginning of each sentence. Other adjustments made by seniors

[8]Nestor is a famous butler in *The Adventures of Tintin* by the Belgian cartoonist Hergé.

Table 6: Modifications performed by seniors.

Speaker ID	Modifications
S1	No modifications
S4	Suppression of “Nestor”, Infinitive form of verbs, Addition of courtesy words
S6	Suppression or move of “Nestor”
S7	Placing “Nestor” at the end of the sentence
S9	Suppression of “Nestor”, Infinitive form of verbs
S11	Suppression of “Nestor”, Addition of courtesy words, Infinitive form of verbs, Courtesy form of verbs

Table 7: Modifications performed by people with visual impairment.

Speaker ID	Modifications
S2	No modifications
S3	No modifications
S5	No modifications
S8	No modifications
S10	Suppression of “Nestor”, Addition of courtesy words

were: addition of courtesy words, moving “Nestor” to the end of the sentence, and use of the interrogative form to ‘request’ and not ‘order’ something to do a task. Table **??** and Table **??** summarize modifications performed by all participants.

6.4 Adaptation of the participants to the context-aware decision

It was highlighted that voice commands would have a lot of contextual information implicit that the system would have to recover. This was addressed by using contextual information (here the location of the user). It is interesting to notice that although there was a large number of repetitions and a significant drift from the syntax, absolutely none of the participants added any term to clarify the goal of the command. For instance, “turn on **the ceiling lamp**”, “close the **kitchen** blinds” etc. were not found in any of the speech utterances whether they were syntactically correct or not. Also, there was no instance of a sentence related to the correction of an order (e.g., “no I meant the kitchen light”). It is unclear, whether this is due to the good performance of the decision when an order was recognised (i.e., no false commands) or because the possibility that the system needed more information did not come into the participant’s mind. In any case, this suggests that the answers of the system perfectly matched the user goal.

6.5 Ownership of the system by the user

Due to the wide variety of variables that might influence the human learning process and the reduced number of participants, the analysis of the familiarisation of the participants with the system was based on what they reported to feel during the experiment. Two topics were discussed with them during the debriefing: (1) the difficulty of use and (2) the ownership of the voice command grammar.

Difficulties felt during the first use No senior reported to have had any difficulty using the system at the beginning. Only one of them indicated she took time to understand she had to repeat orders but she did not consider this repetition constraint as a difficulty but as a parameter of the system. The participants with visual impairment found the time of execution too long (time between the speech utterance and the smart home action). S2 and S5 reported this as **the** difficulty during the first use. Once understood it was necessary to wait, they did not experience any problem.

Acceptance of the voice command grammar No participant said they felt difficulties taking ownership of the system command grammar. Half of participants found the voice commands intuitive (S1, S6, S11, S2, S5, S10). All other seniors (S4, S7 and S9) and one of the visually impaired participants (S8) considered the imperative form of the order sentences as not suitable for the task. S3 found the sentences too childish and would prefer sentences closer to those he uses when communicating with people.

6.6 Users' perception of the system

6.6.1 Satisfaction of vocal interaction

As said before, participants reacted differently when using the voice command to interact with the system. To investigate further on this aspect, they were asked whether they would prefer using voice or manual interaction to interact with the home. Most seniors enjoyed voice interaction (4/6) but they do not consider it necessary and do not wish to have it as the only modality. Voice command would be preferred for difficult tasks (e.g., close or open the blinds) but participants are used pressing the button to control the light and they find it easier. Participants with visual impairment did not answer the question in the same way. They have to adapt their habits to their disability. Three of them would prefer the vocal command over the tactile command. One answered that he prefers the manual command as long as possible but when he would no longer be capable of seeing the switches, he might adopt voice command. The last participant reported he usually touches things to do anything and therefore prefers manual interaction.

6.6.2 Benefits and drawbacks of the SWEET-HOME system

At the end of the session, participants expressed what they think to be the benefits and drawbacks about using the SWEET-HOME system. For seniors, a smart home control system increases the comfort level performing difficult tasks and might allow them to remain independent (ask the

system to do a task rather than asking someone). As they said during the interviews, domestic tasks take time, so they would appreciate delegating several long tasks to the system. The use of a smart-home control system implies two main drawbacks. First, seniors are afraid to end up doing nothing in their home. Second, several seniors said that maybe they would feel they are not able to do something alone. In addition to the fear of becoming dependent on the system, participants with visual impairment are afraid to lose control of the habitat. Due to their visual impairments, they want to keep control and so they need to control their home to feel safe. Their sight problem also implies spending more time on the task, as S5 said : *"la gestion du temps lorsqu'on a un handicap c'est accepter le fait que tout se prolonge dans le temps*" (*"Time management, when you have a disability, is to accept that everything takes more time"*). Thus, the smart-home control system would allow them to be free to do a task when the system is performing another one. From participant feedback, voice interaction and context-aware decision proposed by the SWEET-HOME system partially compensated some of the consequences of their visual impairment. They would feel more independent and more in control in an unknown space because they do not have to look for the objects to control the home (such as the switches). Moreover, the voice interaction limits the necessity to move. Participants of both groups are worried of falling when they move, avoiding risky movements, thanks to a smart-home control system, is an important benefit for the participants. Participants find that the main drawback to the use of the SWEET-HOME system is its slowness that necessitates long pauses before the execution of their orders. Seniors said that another limitation for them is that they are unused to voice interaction.

To detect and analyse voice, the SWEET-HOME system includes microphones placed in the home. During the interview, we asked participants if the presence of microphones was uncomfortable. No participant considered microphones as a drawback for the use of the system. Two participants with visual impairments indicated that the recording should not be accessible out of the home (S3) and that inhabitants should be able to stop the recording (S5). However, the final system would not include recording of raw data.

Last, the participants differently felt the "presence" of the SWEET-HOME system. Seniors felt they are no longer alone by using this system, while participants with visual impairments had the feeling they were alone in the apartment and speaking out loud despite that. This notion of presence is a very hot topic in the companion area (**?**) but it is less investigated in VUI.

6.6.3 Needs for the SWEET-HOME System

Participants had already made some adaptations in order to stay independent. Adaptations made by seniors focused on the layout of their home or the addition of a presence detection system. These modifications are often made so they can feel more confident (due to the fear of falling or insecurity) or to make their home more comfortable. Home modifications made by participants with visual impairment were more diverse. They often augmented the capability of their usual objects (e.g., vocal reader for a computer). They adapted objects in order to continue to use them as they usually did before. They try to have a life as "normal" as possible and then, try to minimize the number of modifications of their habitats. Moreover, they (3/5) share their home with others who do not have a visual impairment so, the modifications should not disrupt the life of other residents. Seniors relate installation of a smart-home control system to a need. They would be ready to install it if they become unable to do usual tasks without it but they estimated that it was not a need at the moment. All participants with visual impairment expressed they wanted to

install the SWEET-HOME system in their home. For this profile of users, the SWEET-HOME system was not an object of comfort but a solution that answers a real need.

7 Discussion

The analysis of the results of the experiment sheds light on the three research questions set in the introduction:

1. Is the ASR performance satisfactory for the application? Does it depend on the user?
2. Is the user able to adapt to the system language?
3. What is the behaviour of the user when interacting with no other feedback than the home automation action?

These questions are discussed with respect to the findings of this study in the following subsections.

7.1 Performances of the speech processing chain

As said in section **??**, 18% of the utterances were not treated by the system. Some voice commands were missed (29) because the detection module sometimes over-segmented the utterances. For instance, some users made a pause between the keyword "Nestor" and the rest of the voice command "Nestor ... turn on the light". As a result, the keyword and the command syntagma were seen by the system as two separate utterances neither of which respecting the grammar. This case occurred 18 times in total in the data. A workaround would be to delay the processing of the keyword so that it could be attached to the next utterance before the voice command recognition. However, due to the limited duration of the experiment it is unclear whether this effect will appear only during the learning period (the necessary time for the people to understand how to pronounce voice commands so that it will be recognized by the system) or whether it will be an installed behaviour. Furthermore, it is likely that for a real implementation and to diminish energy consumption, the keyword detection would be implemented on a chip so that audio processing would be activated only when the system is explicitly called. Overall, the detection module gave acceptable performance for the experiment. After the detection, a filtering stage rejected any utterance longer than 2.2 seconds and shorter than 150 ms as well as those whose SNR was negative. 85 speech utterances, among which 25 were correctly uttered voice commands were discarded. Although it is a high percentage (13% of the utterances), this stage succeeded in filtering out many out of grammar sentences and utterances with too high a level of background noise (or with low energy) to be processed by the ASR. Though this stage could be improved, we argue that it is preferable to favour high precision for some missed true voice commands over having false alarms. Regarding the speech/sound discrimination performance, only 10 noise events were misclassified as speech and 20 speech occurrences were not classified as speech (only 5 were voice commands). Though not perfect, the speech/sound discrimination module was satisfying for the experiment. However, it must be noticed that the experiment was performed in a quiet environment with only one person. A real challenge for the system will be to handle the large amount of sound events and background noise that occur in real homes. Two major trends

exist to handle this problem: 1) noise cancellation when the noise source is known (e.g., TV, radio) (**?**) or 2) source separation techniques which have been recently applied to detection of keywords in a noisy home environment (**?**). The related work in this domain (**????**) shows that although it is still an open problem, there are ways to perform robust ASR in noisy conditions with good localisation of the noise sources.

Overall, the performance of the speech recognition was not very good (43.23% WER) and the recall of the voice command not perfect (59% but 100% precision). Technically speaking this is not satisfying but the participants did not report any difficulty in using the system despite the high repetition rate. Nevertheless, there is clearly room for improvement. Indeed, the ASR system used in this experiment was mono-channel and chosen for its rapidity but further experiments with a more sophisticated ASR approach (acoustic models based on subspace GMM) led to a WER of 10.1% on the same data (**?**). Another way to improve the results would be to build models even more specific to the user. It is known that seniors' voice challenges current ASR systems but it has been shown that well tuned acoustic models can compensate for this issue (**??**).

Another issue related to the ASR was the response time which was 1.5 times the utterance duration (e.g., for a one second long voice command the ASR system took 1.5 second to decode it). All participants were unanimous regarding this aspect of the system. There is again room for improvement as the system was often delayed due to the saturation of the sound event pipeline. A better filtering strategy would avoid this problem. However, this feeling might be due to the amount of failures of the system. In case of failure, the participant waited until he was certain that the order was not caught by the system. A feedback strategy based on light or sound indicating whether the system is processing the request or not might decrease this feeling. In any case, the feedback strategy should be adapted to the abilities of the user (e.g., hard of hearing, low vision) and the context of interaction (e.g., in bed, while washing the dishes, in an emergency situation).

7.2 Acceptance of the restricted grammar

Regarding the grammar there is a clear emergence of two profiles among the participants. The first are the ones who strictly respected the grammar and tried to make the system work. This profile is mostly composed of the participants with visual impairment. Unsurprisingly, they are the group most familiar with ICT and VUI for their daily usage. The second profile frequently diverged from the grammar in terms of syntax (e.g., addition of politeness phrases) and due to their speaking rate, they often paused after pronouncing the keyword before uttering the rest of the command. This group is mostly composed of the seniors. Unsurprisingly, they are the group less familiar with ICT and VUI. A number of remarks were made related to the grammar: the name was not always accepted and the sentences were not found natural enough. This calls for a greater adaptation of the voice command grammar to the person's preferences. Recent advances in the domain based on a short learning period may be a way to provide a more customisable system (**?**).

7.3 Ability of the system to act according to user goals: the role of the context

As reported in section **??**, the decision process made only 2 confusions and 3 false commands. These 5 errors provoked a message through the TTS system indicating the time of the day. After a careful investigation, this was due to a bug in the decision process stage in charge of the identification of the voice command, which was easily corrected. For all the other recognized voice commands, the decision process made no mistake. For instance, no blind was opened in an incorrect room. However, the number of voice commands was low and it is likely that increasing the number of possible commands would challenge the ASR system. In the same way, the keyword 'Nestor' is highly confusable with 'store' (the French word for blinds); however, no confusion was observed in the study. The confusion that could result after an increase of the number of voice commands could be addressed by choosing the grammar so that the phonetised possible sentences are as distant from each other as possible. This technique was successfully used in the design of VUIs for speakers with dysarthria (**?**).

The decision uses both the result of the ASR system and the information about the context to extract the implicit information. For instance, when the user says "open the blinds" the location of the blinds is implicit. The system first recognises the kind of command (here, the action of `opening` the object `blind`) and then uses the context (here the location) to decide which blinds in the home were most probably the user goal. Decision models were available for each voice command related to an object that may exist at different places. For instance, when a voice command such as 'turn on the light' was received, then the decision about the lighting was executed, when it is about the blinds, the decision model about the blind was run and so on. Thus, the voice command had a role of trigger but could in the future be used to pass information on to the decision process. For instance in "turn on the light *dimly*", *dimly* could be used either in lieu of the context or to determine which lamp to use (e.g., the ceiling lamp cannot be dimly turned on). This kind of reasoning could be perfectly represented via logic verification techniques such as the ones implemented in the SWEET-HOME OWL ontology (**?**).

In this experiment, most of the context information was composed of the localisation information which was computed from all the available sensors (speech, infra-red sensors, switches and contact-doors). As presented in section **??**, this localisation was perfect at the time at which the decision was made. However for applications in a real world setting, the method is still to be adapted to multi-users situations.

Since the participants were not willing to perform some activities that were perceived as too tiring for them, the activities were highly reduced and the results about the activity recognition not exploitable. However, in a previous experiment involving 15 naive non aged users, the activity recognition which consisted in recognizing activities among {`Eating, Tidying up, Dressing, Sleeping, Resting, Hygiene, Talking`}, showed an accuracy of 65% using an MLN model (**?**). Though this is far from perfect, the estimation of the activity probability enables the decision to choose the correct action. For instance in a lighting situation in the bedroom where the system had to choose between a high and low intensity, the activity recognition output was: hygiene(0.20), dressing (0.16), sleeping (0.28), and resting (0.17) while the ground truth was tidying up (0.08) in the kitchen. In this example, there is a high uncertainty about the actual activity, but using the activity's probabilities (here hygiene, dressing and tidying up would vote for a high light intensity), the controller did choose high intensity despite the most probable

activity was sleeping. This shows the interest of using contextual information for inferring the user goal.

Apart from disambiguation, the contextual information could also be used to reject some voice commands that are unworkable (e.g., “open the blinds” in a room with no windows) or to elaborate more complex commands related to the routine of the user (e.g., “turn on the small light in the kitchen when I am sleeping”). If the system is going to move toward more natural sentences, one of the main problems will be to interpret the referring expressions. For instance one could say “the ceiling lamp” or “the big light” or even “the low-energy light” to refer to the same lamp. It will then be interesting to explore how a logic representation of the world together with the interpretation of the context could be used to perform the decision making. For instance, in case the system catches the lighting order but the object cannot be interpreted, the system could use the fact that in the morning after sleep, the most usual lamp is the bedside one.

7.4 Users’ behaviour and perception of the communicating system

Despite the modest performance of the speech processing system, most of the participants did not have any difficulty in using the system. This is surprising given the high number of repetitions that were observed (one of them repeated the same command 7 times) and given the length of the experiment. So it seems that participants were lenient towards the failures of the system. This was also observed in other studies (**??**) but not in real system experiments.

The most interesting behaviour was observed during the fourth scenario where participants were not guided. Some participants stayed with the grammar while others added politeness or reformulations despite the three previous scenarios in which only correct voice commands were effective. One interesting observation is that although the system was conceived to generate no affect or personification (i.e., using voice instead of a physical switch), some participants considered the system as a presence and as an entity that really talks and must be respected and some other participants were only using it as a tool (e.g., some visually impaired people had the feeling to talk “alone”). This is surprising behaviour given that apart from the Nestor name there was no humanoid property in the system (the TTS that provided the answer to questions was quite robotic). So the best explanation seems that this feeling was provoked by the very speaker who, by using her voice put herself in a human-like dialogue situation. Some other studies found a similar trend. For instance in (**?**), elderly participants interacting with a spoken dialogue system were showing either ‘factual’ or ‘social’ behaviour. Whether this behaviour is related to age and isolation or to lack of familiarity with voice-based systems is still an open question.

Results Regarding the 3 research questions, the results of the experiment allow the following conclusions:

1. The WER performance was below what it is possible to reach in smart homes and the rate of voice command recognition was 59% meaning that about half of the voice commands were not detected. This rate is highly dependant on the user. However, it is worth noticing that the system was highly precise and thus avoided the well know drawback of systems with too many false alarms. Thus, the performance was not sufficient for the application but we have shown that the ASR performance can be improved with better acoustic modelling. Despite this poor performance, some users were willing to have the system in their home.

2. Regarding the adaptation to the grammar two profiles emerged from the study. Most of the elderly people deviated from the grammar while most of the people with visual impairment did respect it. However, it is likely that this behaviour was due to their familiarity with voice based interactive systems. In any case, most of the people was not satisfied with the grammar and wanted to customise it. Despite this, the participants did not encounter any difficulty using the system and found the commands intuitive. So, overall the system was usable by the participants but the grammar clearly needs to be adapted to the user.

3. The hypothesis of the interaction was that once the user had uttered the voice command, the performed command should be sufficient to provide feedback to the user. The experience showed that though there was not any dialogue and that the aim was to use the voice only to enable hands-free voice commands, some participants put themselves in a dialogue situation. Most participants were annoyed by the lack of feedback of the system, and were even reporting feeling a presence in the home. Thus, it seems that whatever the profile of the user (e.g., 'social' or 'factual') a dialogue management is necessary for a VUI in the home (whether through low level feedback or spoken interaction).

8 Conclusions and open perspectives

This paper presented an experiment with elderly participants and people with visual impairment in a voice-controlled smart home, the SWEET-HOME system. The experiment revealed some weaknesses to address for the automatic speech recognition and the need for a better adaptation to the user and the environment. One important complaint was the rigid grammar used for generating the voice commands. Some people simply wanted a different pattern for efficiency, whilst some others were looking for more natural speech. This seems to be due to the fact that most elderly people actually embodied the system while the other group used it as a tool. This personification is perceivable by the politeness phrases they added to the voice commands and by the fact that they were looking for the source of the synthesized voice while they were alone in the home. There is thus a need for a more flexible grammar for the voice commands. However, it is unclear how the personification needs to be treated given the lack of longitudinal data about voice enabled domestic appliances.

Another identified problem common to both groups was the fact that the system never indicated whether it understood a command or whether the command was completed. This is due to the complete lack of feedback other than the performed actions. Thus, some people were waiting for three to four seconds before realising that the system did not catch their command. Feedback must be provided to the user but studies are needed to investigate whether they are to be text, light or sound based.

Despite all these limitations, the SWEET-HOME system had a positive evaluation. Indeed, the greatest fear of all the participants was a loss of autonomy as well as the fear of falling. These people recognized that the SWEET-HOME system could address these problems. Furthermore, people with visual impairment need to know the state of their own home and have difficulty in managing their usual tasks. For these daily living problems, a voice-controlled smart home would answer these specific needs.

Finally, this experiment must be replicated in the field since users might not have the same

behaviour in their own home as in the living lab. For instance, users might develop a much different grammar than in the experimental smart home. It remains to be seen whether the speech processing performance will be adequate in a real uncontrolled home.

9 Acknowledgments

This work is supported by the Agence Nationale de la Recherche under grant ANR-09-VERS-011. The authors would like to thank the participants who accepted to perform the experiments. Thanks are extended to N. Bonnefond and S. Humblot for their support and to E. Esperança-Rodier for her proof reading of this article.

References

(2013). *The second CHiME Speech Separation and Recognition Challenge: Datasets, tasks and baselines*, Vancouver, Canada.

Aman, F., Aubergé, V., and Vacher, M. (2013a). How affects can perturbe the automatic speech recognition of domotic interactions. In *Workshop on Affective Social Speech Signals*, pages 1–5, Grenoble, France.

Aman, F., Vacher, M., Rossato, S., and Portet, F. (2013b). Analysing the Performance of Automatic Speech Recognition for Ageing Voice: Does it Correlate with Dependency Level? In *4th Workshop on Speech and Language Processing for Assistive Technologies*, pages 9–15, Grenoble, France.

Augusto, J. C. (2009). Past, present and future of ambient intelligence and smart environments. In *ICAART*, pages 11–18.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., and Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan, Part 2, Vol. 87, No. 7, 2004*, 87(2):49–57.

Badii, A. and Boudy, J. (2009). CompanionAble - integrated cognitive assistive & domotic companion robotic systems for ability & security. In *1st Congres of the Société Française des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG'09)*, pages 18–20, Troyes.

Baeckman, L. and Small, B. aud Whlin, A. (2001). *Aging and memory: cognitive and biological perspectives*, chapter Handbook of the Psychology of Aging, pages 349–377. 5th ed. Academic Press, San Diego.

Bailey, P. E. and Henry, J. D. (2008). Growing less empathic with age: Disinhibition of the self-perspective. *Journals of Gerontology: Series B*, 63(4):P219–P226.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. (1977). Gus, a frame-driven dialog system. *Artificial Intelligence*, 8(2):155 – 173.

Bouakaz, S., Vacher, M., Bobillier-Chaumon, M.-E., Aman, F., Bekkadja, S., Portet, F., Guillou, E., Rossato, S., Desserée, E., Traineau, P., Vimon, J.-P., and Chevalier, T. (2014). CIRDO: Smart companion for helping elderly to live at home for longer. *Innovation and Reasearch in BioMedical engineering*, 35(2):101–108.

Branigan, H. P., Pickering, M. J., Pearson, J., and McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355 – 2368.

Callejas, Z. and López-Cózar, R. (2009). Designing smart home interfaces for the elderly. *SIGACCESS Newsletter*, 95:10–16.

Casanueva, I., Christensen, H., Hain, T., and Green, P. (2014). Adaptive speech recognition and dialogue management for users with speech disorders. In *Proceedings of Interspeech 2014*, pages 1033–1037, Singapore.

Cavazza, M., de la Camara, R. S., and Turunen, M. (2010). How was your day?: a companion eca. In *AAMAS*, pages 1629–1630.

Chahuara, P., Portet, F., and Vacher, M. (2011). Fusion of Audio and Temporal Multimodal Data by Spreading Activation for Dweller Localisation in a Smart Home. In *STAMI Series, Space, Time and Ambient Intelligence*, pages 17–21, Barcelona, Spain.

Chahuara, P., Portet, F., and Vacher, M. (2013). Making Context Aware Decision from Uncertain Information in a Smart Home: A Markov Logic Network Approach. In *Ambient Intelligence*, volume 8309 of *Lecture Notes in Computer Science*, pages 78–93, Dublin, Ireland. Springer.

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

Charalampos, D. and Maglogiannis, I. (2008). Enabling human status awareness in assistive environments based on advanced sound and motion data classification. In *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, pages 1:1–1:8.

Christensen, H., Casanueva, I., Cunningham, S., Green, P., and Hain, T. (2013). homeservice: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition. In *4th Workshop on Speech and Language Processing for Assistive Technologies*, pages 29–34.

Clément, S. and Membrado, M. (2010). *Penser les vieillesses. Regards anthropologiques et socioologiques sur l'avancée de l'âge*, chapter Expériences du vieillir : généalogie de la notion de déprise, pages 109–128. Paris : Seli Arslan.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmueller, M., and Maragos, P. (2014). The DIRHA simulated corpus. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 2629–2634, Reykjavik, Iceland.

Cumming, E. and Henry, W. E. (1961). *Growing Old: The Process of Disengagement*. Basic Books, New York.

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

Filho, G. and Moir, T. J. (2010). From science fiction to science fact: a smart-house interface using speech technology and a photo-realistic avatar. *International Journal of Computer Applications in Technology*, 39(8):32–39.

Fleury, A., Vacher, M., Portet, F., Chahuara, P., and Noury, N. (2013). A French corpus of audio and multimodal interactions in a health smart home. *Journal on Multimodal User Interfaces*, 7(1):93–109.

Fozard, J. and Gordont-Salant, S. (2001). *Changes in vision and hearing with aging*, chapter Handbook of the Psychlogy of Aging, pages 241–266. 5th ed. Academic Press, San Diego.

Friedman, D. S., O'Colmain, B. J., Muñoz, B., Tomany, S. C., McCarty, C., de Jong, P. T. V. M., Nemesure, B., Mitchelln, P., Kempen, J., and Congdon, N. (2004). Prevalence of age-related macular degeneration in the united states. *Archives of Ophthalmology*, 122(4):564–572.

Gemmeke, J. F., Ons, B., Tessema, N., Van Hamme, H., Van De Loo, J., De Pauw, G., Daelemans, W., Huyghe, J., Derboven, J., Vuegen, L., Van Den Broeck, B., Karsmakers, P., and Vanrumste, B. (2013). Self-taught assistive vocal interfaces: an overview of the aladin project. In *Interspeech 2013*, pages 2039–2043.

Gödde, F., Möller, S., Engelbrecht, K.-P., Kühnel, C., Schleicher, R., Naumann, A., and Wolters, M. (2008). Study of a speech-based smart home system with older users. In *International Workshop on Intelligent User Interfaces for Ambient Assisted Living*, pages 17–22.

Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M., and Parker, M. (2003). Automatic speech recognition with sparse training data for dysarthric speakers. In *the 8th European Conference on Speech Communication and Technology*, pages 1189–1192, Geneva, Switzerland.

Gregor, P., Newell, A. F., and Zajicek, M. (2002). Designing for dynamic diversity: Interfaces for older people. In *Proceedings of the Fifth International ACM Conference on Assistive Technologies*, Assets '02, pages 151–156.

Hamill, M., Young, V., Boger, J., and Mihailidis, A. (2009). Development of an automated speech recognition interface for personal emergency response systems. *Journal of NeuroEngineering and Rehabilitation*, 6(1):1–26.

Howard, R. and Matheson, J. (1981). Influence diagrams. *Readings on The Principles and Applications of Decision Analysis*, 1 and 2:720.

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

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

Keshavarz, A., Tabar, A. M., and Aghajan, H. (2006). Distributed vision-based reasoning for smart home care. In *ACM SenSys Workshop on Distributed Smart Cameras*, pages 1–5.

Kjeldskov, J. and Skov, M. B. (2007). Studying usability in sitro: Simulating real world phenomena in controlled environments. *International Journal of Human-Computer Interaction*, 22(1&2):7–36.

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

Lecouteux, B., Vacher, M., and Portet, F. (2011). Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions. In *Proc. InterSpeech*, pages 2273–2276, Florence, Italy.

Linarès, G., Nocéra, P., Massonié, D., and Matrouf, D. (2007). The LIA speech recognition system: from 10xRT to 1xRT. In *Proc. TSD'07*, pages 302–308.

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

López-Cózar, R., Callejas, Z., and Montoro, G. (2006). Ds-ucat: A new multimodal dialogue system for an academic application. In *Proc. of Interspeech 2006, Satellite Workshop Dialogue on Dialogue, Multidisciplinary Evaluation of Advanced Speech-Based Interactive Systems*, pages 47–50, Pittsburg, Pennsylvania, USA.

López-Cózar, R. and Callejas, Z. (2010). Multimodal dialogue for ambient intelligence and smart environments. In Nakashima, H., Aghajan, H., and Augusto, J. C., editors, *Handbook of Ambient Intelligence and Smart Environments*, pages 559–579. Springer US.

Mileo, A., Merico, D., and Bisiani, R. (2011). Reasoning support for risk prediction and prevention in independent living. *Theory and Practice of Logic Programming*, 11(2-3):361–395.

Milhorat, P., Istrate, D., Boudy, J., and Chollet, G. (2012). Hands-free speech-sound interactions at home. In *Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012*, pages 1678 –1682.

Moncrieff, S., Venkatesh, S., and West, G. A. W. (2007). Dynamic privacy in a smart house environment. In *IEEE Multimedia and Expo*, pages 2034–2037.

Mueller, P., Sweeney, R., and Baribeau, L. (1984). Acoustic and morphologic study of the senescent voice. *Ear, Nose, and Throat Journal*, 63:71–75.

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

Newel, A. and Gregor, P. (2001). User Sensitive inclusive Design. In *Actes du Colloque Interaction Homme Machine et Assistance*, pages 18–20, Metz, France.

Ogg, J. and Bonvalet, C. (2006). L'état des enquêtes sur l'entraide en europe. Rapport final de recherche CNAF, Mire, Collections de l'INED.

Peetoom, K. K. B., Lexis, M. A. S., Joore, M., Dirksen, C. D., and De Witte, L. P. (2014). Literature review on monitoring technologies and their outcomes in independently living elderly people. *Disability and Rehabilitation: Assistive Technology*, pages 1–24.

Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M. S., and Braga, D. (2012). Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese. In *Proceedings of Advances in Speech and Language Technologies for Iberian Languages (IberSPEECH 2012 Conference)*, pages 139–147, Madrid, Spain.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–190.

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

Portet, F., Vacher, M., Golanski, C., Roux, C., and Meillon, B. (2013). Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17:127–144.

Ravanelli, M. and Omologo, M. (2014). On the selection of the impulse responses for distant-speech recognition based on contaminated speech training. In *Proceedings of Interspeech 2014*, pages 1028–1032, Singapore.

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

Rosow, I. (1974). *Socialization to old age*. University of California Press, Berkeley and Los Angeles, California.

Rotili, R., Principi, E., Squartini, S., and Schuller, B. (2013). A real-time speech enhancement framework in noisy and reverberated acoustic scenarios. *Cognitive Computation*, 5(4):504–516.

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

Ryan, W. and Burk, K. (1974). Perceptual and acoustic correlates in the speech of males. *Journal of Communication Disorders*, 7:181–192.

Sasa, Y. and Auberge, V. (2014). Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the "socio-affective glue". In *SpeechProsody 2014*, Dublin, Irlande.

Schilit, B., Adams, N., and Want, R. (1994). Context-aware computing applications. In *Proceedings of the Workshop on Mobile Computing Systems and Applications*, pages 85–90. IEEE Computer Society.

Sehili, M., Lecouteux, B., Vacher, M., Portet, F., Istrate, D., Dorizzi, B., and Boudy, J. (2012). Sound Environnment Analysis in Smart Home. In *Ambient Intelligence*, volume 7683 of *Lecture Notes in Computer Science*, pages 208–223, Pisa, Italy.

Slavík, P., Němec, V., and Sporka, A. (2005). Speech based user interface for users with special needs. In Matoušek, V., Mautner, P., and Pavelka, T., editors, *Text, Speech and Dialogue*, volume 3658 of *Lecture Notes in Computer Science*, pages 743–743. Springer Berlin / Heidelberg.

Takeda, N., Thomas, G., and Ludlow, C. (2000). Aging effects on motor units in the human thyroarytenoid muscle. *Laryngoscope*, 110:1018–1025.

Vacher, M., Fleury, A., Portet, F., Serignat, J.-F., and Noury, N. (2010). *Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living*, chapter 33, pages 645 – 673. Intech Book.

Vacher, M., Istrate, D., and Serignat, J. (2004). Sound detection and classification through transient models using wavelet coefficient trees. In *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, pages 1171–1174, Vienna, Austria.

Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. (2014a). The Sweet-Home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506, Reykjavik, Iceland.

Vacher, M., Lecouteux, B., Istrate, D., Joubert, T., Portet, F., Sehili, M., and Chahuara, P. (2013). Experimental Evaluation of Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home. In *4th Workshop on Speech and Language Processing for Assistive Technologies*, pages 99–105, Grenoble, France.

Vacher, M., Lecouteux, B., and Portet, F. (2012). Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment. In *Proc. of European Signal Processing Conference (EUSIPCO)*, pages 1663–1667, Bucarest, Romania.

Vacher, M., Lecouteux, B., and Portet, F. (2014b). Multichannel automatic recognition of voice command in a multi-room smart home : an experiment involving seniors and users with visual impairment. In *Proceedings of Interspeech 2014*, pages 1008–1012, Singapore.

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

Vipperla, R., Renals, S., and Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. In *Proc. Interspeech*, pages 2550–2553, Brisbane.

Vipperla, R. C., Wolters, M., Georgila, K., and Renals, S. (2009). Speech input from older users in smart environments: Challenges and perspectives. In *HCI International: Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, number 5615 in Lecture Notes in Computer Science, pages 117–126. Springer.

Wolters, M. K., Georgila, K., Moore, J. D., and MacPherson, S. E. (2009). Being old doesn't mean acting old: How older users interact with spoken dialog systems. *TACCESS*, 2(1).

Wölfel, M. and McDonough, J. W. (2009). *Distant Speech Recognition*. Wiley, New York.

Ziefle, M. and Wilkowska, W. (2010). Technology acceptability for medical assistance. In *pervasivehealth*, pages 1–9.

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