Speech analysis for Ambient Assisted Living : technical and user design of a vocal order system*

Michel Vacher[1], François Portet[1], Benjamin Lecouteux[1], Caroline Golanski[2]

1 Laboratoire d'Informatique de Grenoble UMR 5217,
CNRS / UJF-Grenoble 1 / Grenoble-INP, 38041 Grenoble, France
Tel.: +33 (0)4 76 63 55 73
{Michel.Vacher,Francois.Portet,Benjamin.Lecouteux}@imag.fr

2 MULTICOM, Floralis - UJF Filiale, 6 allée de Bethléem,
38610 Gières, France
Caroline.Golanski@imag.fr

Abstract

Evolution of ICT led to the emergence of smart home. A Smart Home consists in a home equipped with data-processing technology which anticipates the needs of its inhabitant while trying to maintain their comfort and their safety by action on the house and by implementing connections with the outside world. Therefore, smart homes equipped with ambient intelligence technology constitute a promising direction to enable the growing number of elderly to continue to live in their own homes as long as possible. However, th e technological solutions requested by this part of the population have to suit their specific needs and capabilities.

It is obvious that these Smart Houses tend to be equipped with devices whose interfaces are increasingly complex and become difficult to control by the user. The people the most likely to benefit from these new technologies are the people in loss of autonomy such as the disabled people or the elderly which cognitive deficiencies (Alzheimer). Moreover, these people are the less capable of using the complex interfaces due to their handicap or their lack ICT understanding. Thus, it becomes essential to facilitate the daily life and

* Reference of the published paper of this draft:
@article{BookChapterVACHER2013,
chapter = "Speech analysis for Ambient Assisted Living : technical and user design of a vocal order system",
author = "Michel Vacher and François Portet and Benjamin Lecouteux and Caroline Golanski",
title = "Telhealthcare Computing and Engineering: Principles and Design",
publisher = "CRC Press, Taylor and Francis Group, London",
editor = "Fei Hu",
number = "21",
year = "2013",
pages ="607–638",
note = "ISBN: ISBN-978-1-57808-802-7",
keywords = "Ambient Assisted Living (AAL), Home Automation, Audio Analysis, Vocal Order, Safety and Comfort",
}

the access to the whole home automation system through the smart home. The usual tactile interfaces should be supplemented by accessible interfaces, in particular, thanks to a system reactive to the voice ; these interfaces are also useful when the person cannot move easily.
Vocal orders will allow the following functionality: - To ensure an assistance by a traditional or vocal order. - To set up a indirect order regulation for a better energy management. - To reinforce the link with the relatives by the integration of interfaces dedicated and adapted to the person in loss of autonomy. - To ensure more safety by detection of distress situations and when someone is breaking in the house. This chapter will describe the different steps which are needed for the conception of an audio ambient system.
The first step is related to the acceptability and the objection aspects by the end users and we will report a user evaluation assessing the acceptance and the fear of this new technology. The experience aimed at testing three important aspects of speech interaction: voice command, communication with the outside world, home automation system interrupting a person's activity. The experiment was conducted in a smart home with a voice command using a Wizard of OZ technique and gave information of great interest.
The second step is related to a general presentation of the audio sensing technology for ambient assisted living. Different aspect of sound and speech processing will be developed. The applications and challenges will be presented.
The third step is related to speech recognition in the home environment. Automatic Speech Recognition systems (ASR) have reached good performances with close talking microphones (e.g., head-set), but the performances decrease significantly as soon as the microphone is moved away from the mouth of the speaker (e.g., when the microphone is set in the ceiling). This deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise such as TV, radio and devices. This part will present a system of vocal order recognition in distant speech context. This system was evaluated in a dedicated flat thanks to some experiments.
This chapter will then conclude with a discussion on the interest of the speech modality concerning the Ambient Assisted Living.

KEYWORDS: Ambient Assisted Living (AAL), Home Automation, Audio Analysis, Vocal Order, Safety and Comfort.

1 Introduction

Evolution of ICT led to the emergence of the smart home concept which covers all houses equipped with home automation devices, media devices, white and brown devices, central heating, security alarms, etc. that are connected through dedicated communication networks and which are able to function autonomously or to be controlled via a private centralised controller. These smart homes are characterised with a high degree of human to machine (home) interaction and/or a proactive management of the home. Smart homes objectives are numerous and include security improvement, comfort, energy efficiency, remote home control, monitoring and daily assisted living.

An important push to the development of smart homes is the emergence of the Ambient Assisted Living (AAL) objective which aims at addressing the challenges of the elderly demographic increase thought ICT solutions. Indeed, the continuous rise of the elderly population will lead to a situation in which the youngest will not be sufficiently numerous to care for their seniors. Despite many progresses in medicine, ageing is correlated with higher risk of loosing independence, due to physical or cognitive decline or disease, than the rest of the population. AAL solutions are being developed in robotics, home automation, cognitive science, computer network, etc. to compensate for the loss of autonomy of our elders and to enable them to live as unhampered as possible in their own home.

In the context of AAL, the primary tasks of the smart homes are to monitor the inhabitant and provide them with assistance when needed. The advantages of this technology is two-folds:

- when well designed, it permits increased self-reliance and control over their lives, and greater possibilities for social interaction, all that having a major impact on their sense of well being and health (Rodin, 1986) ;
- it supports the work of professional carers and reassure relatives and close friends.

The kind of support that is provided by these smart homes is function of the kind of autonomy loss of the person. It ranges from automatic light control to fall detection or prevention. For instance, when someone is waking up at night to go to the toilet, sensors can detect this situation and provide personalised light path to guide the person. If the person is blind, voice command can be used to control her environment, if the person has some memory loss, the system can warn about security issues (e.g., cooker kept on for a long time), ...

This technology has a great potential to ease the life of the elderly population however, the technological solutions requested by this part of the population have to suit their specific needs and capabilities. Indeed, these people are the less capable of using the complex interfaces due to their handicap or their lack of ICT understanding. Thus, it becomes essential to facilitate the daily life and the access to the whole home automation system through the smart home. They should keep control of their environment and not undergo the acts of the house. To make this possible, the usual tactile interfaces should be supplemented by accessible interfaces, in particular, thanks to a system reactive to the voice; these interfaces are also useful when the person cannot move easily. To achieve this vision, we set up the SWEET-HOME project[1] that aims at designing a new smart home system based on home automation technology and on audio technology..

The Sweet-Home project (Vacher et al., 2011a) is a French national supported research project, that is made up of researchers and engineers from two academic labs: the Laboratory of Informatics of Grenoble[2] (specialised in speech processing, smart home

[1] `sweet-home.imag.fr`
[2] `www.liglab.fr`

design and evaluation) and the Esigetel[3] (specialised in audio technology) and from three companies: Theoris (real-time system development and integration), Camera Contact (diffusion of adapted services for maintenance at home) and Technosens (videophone equipment for the elderly). The three following functionalities will be ensured by the system:

- to provide assistance via *natural man-machine interaction* (voice and tactile command),
- to facilitate *social contact*, and,
- to provide *security reassurance* by detecting situations of distress.

If these aims are achieved, then the person will be able to pilot, from anywhere in the house, her environment at any time in the most natural way possible.

The project tries to make use of already standardised technologies and applications rather than building communication buses and purpose designed material from scratch. As emphasized in (Mäyrä et al., 2006), standards ensure compatibility between devices and ease the maintenance as well as orient the smart home design toward cheaper solutions. The interoperability of ubiquitous computing elements is a well known challenge to address (Edwards and Grinter, 2001). Another example of this strategy is that Sweet-Home includes already designed systems such as the e-lio [4] or Visage [5] systems which will make it easier for the user to connect with their relative, grocer or caregiver. We believe this strategy is the most realistic one given the large spectrum of skills that are required to build a complete smart home system. Moreover, in order for the user to be in full control of the system and also in order to adapt to the users' preferences, three ways of commanding the system are possible: voice order, tablet computer or classical tactile interface (e.g. switch).

This chapter describes the different steps which are necessary to the conception of a audio-based smart home system. After a brief overview of the literature on audio sensing technology in smart home in Section 2, the way of taking into account the wishes, expectations and fears of the users is presented in Section 3 and a user study assessing the acceptability of this technology in the elderly population is described Section 4. In Section 5, the technical aspect of the SWEET-HOME project and the audio technology is detailed. Then, in Section 6, the chapter presents a series of experiments undertaken to improve speech recognition in smart homes. The chapter finishes with a short conclusion and list some promising research directions.

[3] www.esigetel.fr
[4] www.technosens.fr
[5] camera-contact.com

2 Audio sensing technology in Smart Home

In the context of smart homes, audio analysis can be divided into two main branches:

1. speech recognition for voice command and dialogue, and
2. sound recognition to identify the interaction of the individual with her environment (e.g., closing a door) or with a device (e.g., vacuum).

Speech recognition is a very old research area which, despite many progresses in close talking situations, must address numerous challenges in noisy distance speech context (e.g., in a home). Sound recognition in domestic areas (washing machine, keys, door...) has been considered as an interesting research area in smart home since the 2000's. Though they are not directly useful to command a smart home, when well detected, they can support health monitoring (e.g., scream, cough, snoring...) and can serve for disambiguation and activity recognition purpose (e.g., dishes manipulation, water, glass break...). In this section, a brief state of the art in this field is presented, then the challenges that are still to overcome in the sound analysis for home care domain are introduced.

2.1 Audio and speech analysis

Regarding the identification of sounds, some projects aim at determining the general state of health of the person living in an intelligent building through the recognition of activities or the detection of distress situations. Noise analysis was used to quantify the use of water (drinking and sanitation) (Ibartz et al., 2008) but it is rarely used for the detection of distress (Istrate et al., 2006). Nevertheless, Litvak et al. (Litvak et al., 2008) used microphones and an accelerometer to detect falls in an flat, while Popescu et al. (Popescu et al., 2008) used two microphones for the same purpose. Outside the context of distress detection, Chen et al. (Chen et al., 2005) identified various activities in a bathroom using HMM on cepstral coefficients mel-frequency (Mel-Frequency Cepstral Coefficients, MFCC). Sound event recognition has many applications in robotic (Geiger et al., 2011) and a robust recognition in noisy condition is known to be a difficult challenge (Leng et al., 2011).

The speech recorded by a microphone is characterized by a spectrum with frequencies from 50 Hz to 8 kHz. This spectrum depends on the speaker and on the uttered words. The words making up a sentence can be decomposed into a sequence of phonemes that are basic sounds. Each language has its own phonemes, 37 phonemes in the case of French. Each phoneme is affected by the preceding and by the following phoneme, this phenomenon is called coarticulation. Speech production requires airflow from the lungs (respiration) to be phonated through the vocal folds of the larynx (phonation) and resonated in the vocal cavities shaped by the jaw, soft palate, lips, tongue and other articulators (articulation). It is clear that the realization of a speech recognition system

depends on the spoken language and should take into account the variability introduced by the various speakers. Moreover, the same sentence will be pronounced differently by the same speaker according to her fatigue or her mood.

In Automatic Speech Recognition (ASR), many methods using varied acoustic parameters were explored and the state of the art of this technology is strongly linked to the automatic learning of probabilistic models (Hidden Markov Models) for phonem modelling(Rabiner, 1989). Most speech recognizers are made up of three interconnected modules:

1. an acoustic stage able to recognize the most probable phonem sequence in lattice form,

2. an acoustic dictionary to associate each word to its phonem decomposition,

3. a language model (LM) that is a collection of constraints on the sequence of words. In order to realize large vocabulary recognizer, the language model must be related to an entire language; for particular applications, the language model can be specific with a small vocabulary dedicated to the application.

Besides the commercially available systems, there are many ASR that are available as open software such as Sphinx from the CMU (Seymore et al., 1998) or Speeral from the LIA (Nocera et al., 2002), HTK from the Cambridge University, RWTH ASR from the Aachen University and many others.

Recent developments have produced significant results and enabled the ASR to be a component of many industrial products, but there are still many challenges to make this feature available for Smart Homes. For instance, ASR systems achieved good performance when the microphone is placed near the speaker (e.g., headset), but the performance degrades rapidly when the microphone is placed at several meters (e.g., in the ceiling) (Wölfel and McDonough, 2009). This is due to different phenomena such as the presence of noise background and reverberation (Vacher et al., 2008). These issues must be considered in the context of home assistance, some solutions will be presented in section 6.

2.2 Challenges to overcome

The audio analysis (of sounds and speech) has a great potential for monitoring the person, disability compensation, assistance, and security enhancement. It can also be useful for improving and facilitating communication of the person with the outside. However, as previous experiments confirmed (Vacher et al., 2011b), this technology (recognition of speech, dialogue, speech synthesis, sound detection) must take into account the very difficult environment of the home. In this section we will present some possible applications and the main challenges to address.

2.2.1 Extraction of audio in a noisy environment

In real conditions, the processing of the audio channel is often disturbed by: 1) the presence of unwanted background noise (TV, various devices, traffic), 2) the acoustics of the room and; 3) the speaker's position and orientation with respect to the microphone. To study the problems encountered in a real environment, we conducted an experiment in a dedicated smart home. This experiment (Vacher et al., 2011b) involved 15 people who carried out the activities of daily living for about one hour each. The sounds generated were registered under realistic conditions by microphones set up in the ceiling. In the database of sounds collected, the noise level varies between 5 and 15 dB, which is very far from the ideal studio conditions. The signal level varied depending on the participant's position in the room. Indeed the participants were free to move while they were talking as far as they wanted from the microphones (set in the ceiling). The presence of very large windows facing each other led to alterations in the signal (reverberation). What was the most striking problem was the presence of simultaneous signals, due to sources parasites such as TV set. To overcome these problems, signal processing techniques taking into account simultaneously all or some of the microphones should be considered (blind source separation, principal component analysis, beam-forming...).

Blind Source Separation (BSS) is a classical method frequently used for music and speech processing in the case of additive mixing (Benabderamane et al., 2011; Sarmiento et al., 2011; N. Tran, 2011). These studies are more and more applying to speech analysis in various environment (CHiME, 2011). First studies showed that in our recording conditions, convolutive mixing is operated and then specific algorithms must be adapted.

2.2.2 Monitoring of activity and health

Audio analysis can be quite useful for the monitoring of the person's activity and the assessment of some decline. For instance, an application might consist of recognising the household appliances or water usage to assess how the person uses her environment to carry her its activities. Health related symptoms such as coughing, scraping throats and sniffles can also be detected to gather information about a potential health problem. For home automation applications, detecting sound sources can be useful to act depending on the location of the person to care for (lighting that follows the person at night...). In activity recognition, speech detection plays a fundamental role in the identification of communication phases (Portet et al., 2009). Social isolation can then be detected based on speech recognition, phone usage and other textual-based communication means. A more ambitious application would be to identify non-verbal communication to assess the welfare or suffering of a person with dementia.

2.2.3 Evaluation of the ability to communicate

The speech recognition could play an important role in monitoring people with dementia. Indeed, one of the most tragic symptoms of Alzheimer's disease is the progressive loss of

vocabulary and communication skills. Constant monitoring of activity of the person could allow the detection of important stages in the evolution of dementia. Audio analysis is the only modality that offers the possibility of monitoring the automatic loss of vocabulary, the decay times of speaking, the isolation in conversation, etc.. The changes can be very slow and difficult to detect by the carers.

2.2.4 Voice interface for compensation and comfort

The most direct application of voice interface is to provide the ability to command verbally the smart home to enable the physically disabled persons (e.g., person in wheel-chair, blinds, etc) to keep control of their environment. With such users, tactile interface might be very difficult to operate, while non-physical interface such as voice command free them from moving or seeing. For instance, when an elderly person wake up at night, she can ask the system to turn on the light rather that searching blindly for the switch.

2.2.5 Detection of distress situations

Identifying the sounds of everyday life can be particularly useful for detecting distress situations in which the person might be. For example, the detection of a glass breaking is commonly used in alarm systems. In addition, if the person remains conscious but can not move, for example after a fall, the system offers the opportunity to call for help by the voice. Moreover, taking the emotional aspect into account is an other important aspect as pointed out by gerontologists (Cornet et al., 2007) and physicians. It could be very profitable to evaluate the emotion level in speech: in case the speaker feel a great emotion, the system could take this information into account before making its decision. The presence of emotion is as important as the recognition of the sentence itself because this is a good indication that the person has a problem to be solved. We must also consider the fact that the recognition of emotional speech is more difficult than classical speech recognition.

2.2.6 Challenges in the recognition of sounds of everyday life

The recognition of everyday life sounds is a relatively new field, and acoustic parameters as well as classifiers are not standardized (Dufaux, 2001; Cowling, 2004; Istrate et al., 2006; Fleury et al., 2010; Tran and Li, 2009). In the literature, the methods depend, in most cases, on techniques that learns probabilistic sound models from corpora. However, the experiments that we conducted in the HIS (Vacher et al., 2011b) showed that there is a great diversity of classes of sounds. Moreover, in most cases, the total duration of each class is very small, because the class includes very few elements. These characteristics make the usage of classical learning methods difficult (i.e., large number of classes with a very imbalanced number of examples in each class). Moreover, theses corpora are usually annotated with the source of the sounds (e.g., chair displacement) while a more natural

classification would be to use the inner acoustic characteristics (e.g., rubbing). Thus, acquisition of realistic and well annotated corpora poses real challenges in the smart home domain.

To address this problem, a hierarchical classification method based on the intrinsic characteristics of the signal (fundamental frequency, periodicity, envelope shape...) could be a solution for improve the system. Another option would be to remove ambiguities by using other sources of information available in the house to determine the context of occurrence of the sound. A intelligent supervision system could then use this information in order to make an appropriate decision.

2.2.7 Speech recognition adapted to the speaker

Several experiments carried out in automatic speech recognition showed degradation of performances with 'atypical' people such as children or the elderly (Wilpon and Jacobsen, 1996; Vipperla et al., 2008; Gerosa et al., 2009). Other studies (Gorham-Rowan and Laures-Gore, 2006; Vipperla et al., 2010) emphasized the effects of ageing on speech production and the implications this has on speech recognition. The elderly speakers are characterized by a trembling of the voice, hesitations, the production of inaccurate consonants, a breaking voice, and a slower articulation. The speech recognition from elderly voice is an under explored area (Dugheanu, 2011).

2.2.8 Privacy and acceptability

Given the increased trend in fitting houses with more and more sophisticated ICT devices, the question of privacy in ones own home is emerging (van Hoof et al., 2007; Sharkey and Sharkey, 2012). As any other technologies, speech recognition must respect the privacy of the speaker. Therefore, the language model must be adapted to the application and should not allow to recognize phrases whose meaning is not essential to the application; a keyword recognition system respects this constraint. Moreover, it is an open question whether the system can store the audio signals or not (and for how long).

Regarding the acceptability aspect, a system will be far better accepted if it is a useful daily living assistant rather than a occasional one (e.g., fall detector). A general system covering surveillance, home automation and detection of distress, would be more easily accepted than scattered one-purpose ICT devices. Acceptability is key aspect of the successful development of smart homes. This is why the user centred approach for the design is presented in the next section. In accordance with this method, we conducted a specific study on the acceptability of a voice interface for older people. This particular study and its results are presented in Section 4.

3 User centred design

The user centred designed applied in Sweet-Homefollows the diagram given in Figure 1. The first step consists in drawing up a set of requirements based on our expertise in smart environments, our expertise on health-care and social equipments for the elderly and on user surveys from the literature (Edwards and Grinter, 2001; Koskela and Väänänen-Vainio-Mattila, 2004; Demiris et al., 2004; Kang et al., 2006; Mäyrä et al., 2006; Fugger et al., 2007; Rialle et al., 2008; Callejas and López-Cózar, 2009). This made it possible to make a first list of specifications of the system which was used to design its functionalities and methods of interaction. The Wizard of Oz (WOZ)[6] step consists in confronting the potential users with a system that they believe to be automatic, but which is actually being operated by an experimenter (Jambon et al., 2010). This provides feedback from the users in a realistic situation. The feedback and suggestions are then incorporated into the design. After the WOZ, the various features of the Sweet-Home system are being developed independently in a more machine centred way (i.e. *making it work*) but including human users in the loop as much as possible (i.e., *making it easy to use*). When developed, all the functionalities will be integrated together with the home automation environment and this real system will be again tested with the targeted users and adapted and corrected if necessary. The section 4 reports the results of the WOZ experiment.

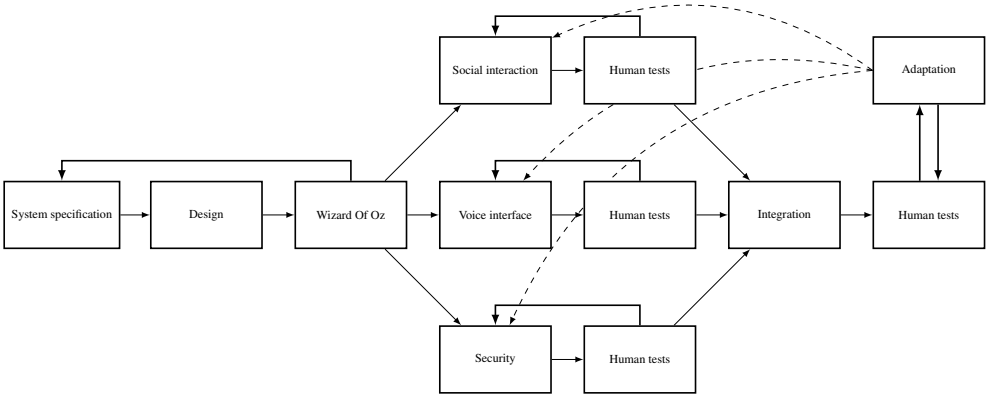

Figure 1: Scheme of a user centred design

4 Acceptability of vocal orders

In our project, the targeted users are people who are elderly but still active and autonomous. The rationale behind this choice is that a home automation system is expensive, so it would be much more profitable if it can be used to accompany daily life rather than

[6] In Wizard of Oz experiments, participants interact with a system that they believe to be automatic, but that is actually remotely controlled by an experimenter — the wizard.

only when the need for assistance appears. Moreover, if the user's situation changes (e.g., wheelchair, cognitive decline), the system could be adapted and specific assisting technologies could be 'plugged-in' to adapt the environment to the user and not to suddenly impose a completely new way of life to the user by fitting her house with ICT. This kind of population contains frail individuals but that does not imply they are not self-governing. Thus, the design of new daily living support technologies must take into account studies that have shown that the reduction of sense of control in the elderly population may have a significant adverse effect on their health (Rodin, 1986). To find out what the needs of this target population are, we conducted a user evaluation assessing the acceptance and the fear of the audio technology (Portet et al., 2011). The experience aimed at testing three important aspects of speech interaction: voice command, communication with the outside world, home automation system interrupting a persons activity. The experiment was conducted in a smart home with a voice command using a Wizard of OZ experiment with interviews of elderly people and their relatives in co-discovery. After a brief outlook of the literature, the experimental design is described. The results of this experiment are then summarised and discussed.

4.1 overview of user studies in audio smart home projects

Many studies have been conducted in different countries to define the needs of elderly concerning a smart home system able to help them in their daily life (Koskela and Väänänen-Vainio-Mattila, 2004; Mäyrä et al., 2006; Demiris et al., 2004; Kang et al., 2006; Callejas and López-Cózar, 2009; Ziefle and Wilkowska, 2010). These studies concern systems that provide support in three main areas: Health monitoring, Security and Comfort. *Health* oriented systems are those which monitor the status of the person (e.g., weight, heart rate, activity) via physiological sensors, movement detector, videos, etc.; *Security* oriented systems provide distress or hazardous situation detection, for instance, fall detection, smoke detection, intrusion detection, etc.; and *Comfort* oriented systems based on classical home automation allowing people to manage home appliance in an easy way.

A number of quality measures were identified in these studies that fall under two major categories: acceptability (Usability, Affordability) and trustworthiness (Safety, Security, Reliability, Privacy). However, as pointed out in (Augusto, 2009), it is difficult to define an unique criterion about user acceptability given the diversity of users and applications considered. Furthermore, some studies used quantitative evaluations others qualitative ones with large samples of persons or small focus groups composed of young or elderly persons. Therefore, these studies are difficult to compare.

Regarding the experimental setting, most of the studies which included a prototype were conducted in temporary spaces fitted with sensors or in a real in-lab flat built for the laboratory. Very few experiments have actually been conducted within the persons' own homes with the notable exception of (Mäyrä et al., 2006). In fact, the smart home domain is still recent and to the best of our knowledge there is no standard procedure that has emerged. Despite the diversity in experimental settings of the studies reported

in the literature, aims, criteria, targeted users and technologies, results of these studies show convergence to some frequently expressed issues : Security, Health Monitoring, Pro-activity (e.g., automatically turning on the light when it is too uncomfortably dark), Usability and respect of Privacy.

One common outcome of all these studies is that, whatever the technology being considered, no smart home application is going to be successful if the intended users are not included in the design of the system (Mäyrä et al., 2006; Fugger et al., 2007; Callejas and López-Cózar, 2009; Augusto, 2009). Acceptability is the key factor to integrating new technologies in homes, particularly when the users are elderly or low ICT educated persons.

4.2 Experimental set-up

The experiment consisted in Wizard of Oz phases and interview phases in the smart home. Seniors, their relatives and professional carers were interviewed and were interacting with the system during about 45 minutes. The interviews were semi-directive (open questions) and held in co-discovery (participants were always grouped by couple). The functionalities of the system to assess were presented to the participants one after another and each participant had time to discover them before being questioned and before the WOZ interaction. The WOZ interaction consisted mainly in the control of the environment. For instance, if the participant said *"close the blind"*, the blind were closed remotely.

The participants consisted of 18 persons from the Grenoble area. Among them, 8 were in the elderly group, 7 in the relatives group (composed of mature children, grandchildren or friends). The mean age of the elderly group was 79.0 (SD=6.0), and 5 out of 8 were women. These persons were single and perfectly autonomous. The frequency of visit by their relative was from once a week to everyday. The mean age of the relative group was 41.0 (SD=19.5), and 5 out of 7 were women. Their relationship with their elder partner varies but they were chiefly grandchildren (4/7). In order to acquire another view about the interest and acceptability of the project system, 3 professional caregivers were also recruited to participate in the experiment. This group was composed of 2 nurses and one professional elderly assistant. These people were mainly recruited to give a general point of view about how elderly persons and their relatives could accept the system and how such a system could also facilitate their daily work.

Each test was composed of an interviewer, a wizard and a couple of one elderly person with a relative (except for one senior who was alone). As shown in Figure 2, the participants and the interviewer were inside the smart home during the whole test except during some parts of the scenarios during which the relative moved to another room (e.g., video-conferencing).

The co-discovery approach was chosen to reassure the senior about the experimental context (new environment, experimenter, etc.) thanks to the presence of their relative. Moreover, it eased the projection of both participants into the new system because they could exchange points of view. Of course, the relationship between the two people can

Figure 2: Picture of a typical interview within the smart home

also influence the experiment (e.g., a grand mother who would not like to expose her weakness to her grand son) that is why some short periods were planned during which the participants were interrogated separately.

The aim of this study was to assess the *acceptability* of the system. But, there is no standard definition about user acceptability in this domain (Augusto, 2009). Indeed, users can reject a new technology because it is too complex, too intrusive, not natural, does not fit their education or religion, does not meet their needs, etc. Thus, most of the experiment was conducted to find out whether the potential users would appreciate the new functionalities brought by the system (e.g., *'Do you appreciate making the system operate using your voice? Why?'*, *'Do you find this natural?'*). Moreover, in order to guide the development of the system, aspects of *usefulness*, *usability*, *personalisation* (how one wants to speak to the house), *interactiveness* (interaction modalities such as voice, tactile or remote control), *proactiveness* (when the system decides to act without human intervention), *intrusiveness* (disruption in the middle of an activity), *social interaction*, and *security* were investigated.

The first phase of the experiment was about the **voice command** aspect of the project. Both the senior and her relative were present in the room. The senior was asked to control blinds, lights and the coffee machine using her voice without any recommendation about how to do it. This consisted in talking *"to the home"*. The vocal order given was followed by the proper action in the home operated by the hidden wizard. For each participant an incomprehension of the system was simulated. Then, questions regarding the naturalness and easiness of the voice command were asked to both persons. Finally, the senior stayed in the smart home with the interviewer while the relative was taken to another room and questions regarding the preferred form of interaction were asked to both separately. For instance, "Do you prefer to talk to a remote control or to the house in general?" "Would you rather use the 'vous' form (formal) or the 'tu' form (familiar) when uttering an order?"

The second phase consisted in using technology for **communication with the outside** such as video conferencing. The senior was left alone in the smart home watching a TV program, when the program interrupts itself to let the face of the relative appear on the screen so that they can start a conversation. After the conversation, the senior was rejoined by her relative and the interviewer. Questions were asked about their own preferences.

The third phase focused on **system interruption**. The couple and the interviewer were discussing in the smart home when the system interrupted them via a pre-recorded voice played through the speakers, calling for a door to be closed or the cooker to be turned off. After this, questions related to whether being interrupted by the system was acceptable or not and how the interruption should happen were asked. Also, the problem of security in general and how such system could enhance security was discussed with the couple.

4.3 Outcomes of the study

From the results of the study it appears that seniors preferred mostly the voice command for the blinds and light (6/8), the system interventions about safety issues (4/8) and the video-conferencing (1/8). Relatives mostly liked the voice command for the blinds and light (5/7), the system interventions about safety issues (2/7) and the video-conferencing (2/7). Carers mostly preferred the system interventions about safety issues (3/3) and the voice command (2/3). The voice command is the preferred feature of the system overall along with the interruptions about security issues. This confirms that a smart home fitted with speech processing technology is a promising technology that should be accepted by the elderly population.

Most of the participants found the voice interface natural. They also had tendency to prefer or to accept the 'tu' form (informal in French) to communicate with the system given this system would be their property. We are not aware of any study investigating this important aspect of acceptability (related to this is (Gödde et al., 2008) who emphasized that elder Germans tend to utter longer and politer commands than their younger fellow countrymen). It is interesting to note that, in our study, the 'key-word' form for commands is highly accepted (rather than the sentence based command). This would enable the system avoiding many of the current bottlenecks in speech recognition (e.g., ambiguity, complete sentence detection, etc.). This highly simplifies the integration of such technology in smart home given that small vocabulary systems are generally performing better in real world applications than large vocabulary ones. Regarding the system communication, half of the seniors would prefer a system with a female voice, one would rather hear a male voice but for the others, this did not matter. They were all unanimous about the fact that the voice system must be natural and not synthetic. This is in line with the findings of a study (Lines and Hone, 2006) investigating the preferences of 32 seniors (over 65) between synthetic and natural voice in different noise conditions. More than 93% preferred the natural voice.

As in other related studies (Callejas and López-Cózar, 2009), all participants found a

strong interest in the voice interaction system. It is strongly preferred over a tactile system (or touch-screen) which would necessitate being physically available at the place the command is to be found or would imply to constantly know where the remote controller is. This is in line with other studies concerning personal emergency response systems which showed that push-button based control is not adapted to this population (Hamill et al., 2009). In other user studies (Callejas and López-Cózar, 2009), they observed that even if the system is sometimes wrong in interpreting orders, most of the people are willing to continue using it. In (Callejas and López-Cózar, 2009), a survey conducted among 200 intended users (50 to 80 years old) of a smart home for the elderly showed that people would tolerate some demands for repetitions in cases when the voice interface does not understand. However, for a reliable assessment of this notion of tolerance to repetition and the way the system provides solutions, it would have been necessary to place the subjects in real conditions for several days so that they were faced with several situations of system misunderstanding.

Although the system was well received, it turned out that some functionalities provoked strong objections among the participants. The main fear of the elderly and relatives is the system failure. Another main concern about the system is the fact that too much assistance would increase the dependence of the person by pushing her toward inactivity. Regarding the carers, they expressed the concern that such system would tend to gradually replace some of their visits and end up in making the seniors even more isolated. Most of these fears can be addressed by a good design of the system. For instance, home automation systems can include several interaction modalities so that when one does not work, others can be used (e.g., voice recognition and classical switches). However, fear about a decrease in autonomy due to a system that can do everything is a subtle one. A system designed for active people in order to improve comfort, security and save time (Koskela and Väänänen-Vainio-Mattila, 2004) may not be adapted to healthy but aged persons. For instance, saving time might no longer be a requirement when the person is retired. Another potential problem is the feeling of intrusion when the system suddenly interrupts the person to remind them an appointment. All groups (seniors, relatives, carers) found the interruption too intrusive and they recommended to warn the person using a short piece of music or sound in order to reassure the person. Indeed, without warning, the impression could be given that someone is entering the house while the person thought she was alone and thus make her panic before she realise the voice is coming from the home. The information delivered in front of someone else was not questioned but may also be a privacy issue (the system should 'know' that the person is alone before delivering information or is using the telephone).

While the proposed system can bring more comfort and autonomy to daily life by providing an easy interaction with the home automation elements, the majority of the participants insisted on the security aspects. For instance, the voice interface would be of great use in case of falls. The elderly and their relatives have particularly appreciated that the system spares the elderly actions that can be dangerous (running to get the telephone handset, finding the switch in the middle of the night to turn on the light) and alerts them of dangerous situations (door opened, gas on, etc.). This trend is confirmed in almost all user evaluations involving elderly (Rantz et al., 2008; Callejas and López-Cózar, 2009;

Rialle et al., 2008) and by the dramatic number of research teams and companies working on fall detection (Noury et al., 2008). Thus, smart homes for the elderly would be much more accepted if they contain features that can reassure them regarding security more than any other features whatever their initial condition and origin in developed countries are.

Overall, the participants mainly stressed the interest of voice command and how this could improve security, autonomy and, to a smaller extend, could fight loneliness. However, they were very careful about privacy and clearly showed that they were very cautious of not accepting systems that would push them into a dependent situation. They want to keep control. Although only a small sample of seniors and relatives in healthy condition was recruited, this qualitative study confirmed the interest of voice-based technology for smart home and uncovered some pitfalls to avoid in its design. For a more detailed description of this user study the reader is refereed to (Portet et al., 2011).

5 Architecture of a audio-based smart home system

As many smart home systems, the Sweet-Home solution contains four mains components:

1. a central analysis and decision unit that interpret orders and situations and make decisions,

2. a range of sensors that perceive events in the environment,

3. a range of actuators which permit to act on the environment,

4. a home automation communication infrastructure that makes message exchanges between the above components possible.

The Sweet-Home architecture is described in Figure 3. The input of the Sweet-Home system is composed of the information from the home automation system transmitted via the local network and of the information from the microphones transmitted through radio frequency channels. An audio analysis module is in charge of the sound recording and the recognition of speech and sounds. The decision stage uses the voice orders extracted from the speech recognition flow as well as the sounds of daily living that give contextual information. The decision stage of the Sweet-Home system will implemented by an intelligent controller which will analysis all the transmitted to interpret them and execute the relevant actions.

The experimental flat DOMUS which was used to test and train the different stages of the smart home system is presented below. The section ends with a detailed presentation of the sound analysis module.

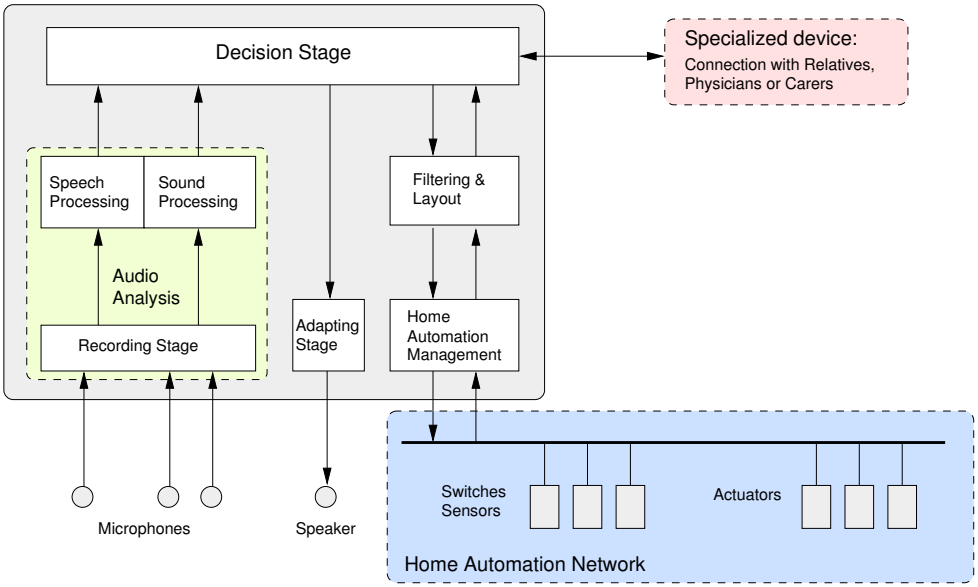

Figure 3: General architecture of the Sweet-Home system.

5.1 The experimental flat DOMUS

The DOMUS smart home was designed by the Multicom[7] team of the Laboratory of Informatics of Grenoble (LIG) to observe users' activities interacting with the ambient intelligence of the environment. Figure 4 shows the details of the flat. It is a thirty square meters suite flat including a bathroom, a kitchen, a bedroom and a study, all equipped with sensors and actuators so that it is possible to act on the sensory ambiance, depending on the context and the user's habits. The flat is fully usable and can accommodate a dweller for several days. The technical architecture of DOMUS is based on the KNX bus system (KoNneX), a worldwide ISO standard (ISO/IEC 14543) for home and building control. More than 150 sensors, actuators and information providers are managed in the flat (e.g., lighting, shutters, security systems, energy management, heating, etc.).

The flat has also been equipped with 7 radio microphones, Sennheiser eW-300-G2, type ME-4, set into the ceiling for real-time sound recording. Each microphone transmitter is connected to a specific receiver with its own High Frequency band (between 656.450 MHz and 661.850 MHz), each receiver output is then wired to an input of the multichannel PCI-6220 acquisition board (National Instruments).

[7] http://multicom.imag.fr/

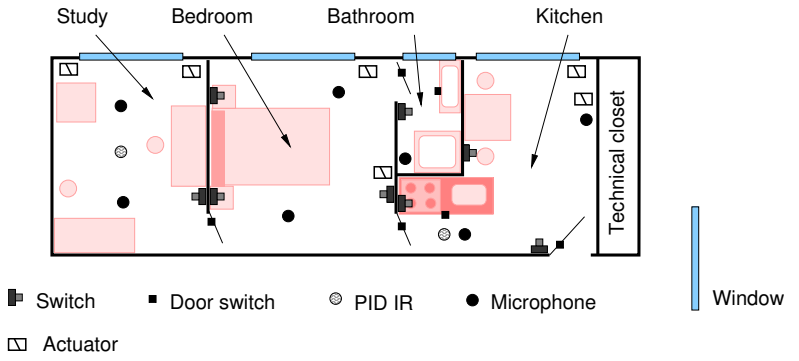

Figure 4: Layout of the DOMUS Smart Home and position of sensors and actuators.

5.2 The audio processing module

The sound analysis module of the Sweet-Home system is based on the AUDITHISarchitecture. The AUDITHISsystem is able to analyse audio events in a smart home in real-time. Speech and sounds are recorded from one to eight microphones placed at different location in the flat. The general organization of the system is depicted in Figure 5.

AUDITHISis running on a Debian 6.0 GNU/Linux operating system (2.6.32 Kernel) and is made of 3 independent processes synchronized through a file exchange protocol:

1. the AUDITHIS**kernel**, that is the central part of the system and will be described later in this section.

2. the **Automatic Speech Recognizer** is described in section 6. This system uses a phoneme representation of the speech signal. Rabiner's published the scaling algorithm for the Forward-Backward method of training of Hidden Markov Model recognizers (Rabiner, 1989) and nowadays, modern general-purpose speech recognition systems are generally based on HMMs as far as the phonemes are concerned. Language Models (LM) were then introduced. A LM is a collection of constraints on the sequence of words acceptable on a given language and may be adapted to a particular application. According to theses choices, the specificities of each recognizer are related to its adaptation to a unique speaker or to a large variety of speakers, and to its capacities of accepting continuous speech, and small or large vocabularies.

3. the **Sound of daily living Classifier**. Two different methods are available to classify everyday life sounds: Gaussian Mixture Models (GMM) or Hidden Markov Model (HMM). These methods give comparable results, the HMM classifier gives the best results in noise free conditions, but the GMM classifier is more robust the SNR is under +10 dB (Vacher et al., 2007). The models were trained with our corpus containing the eight classes of everyday life sounds, using LFCC features; this

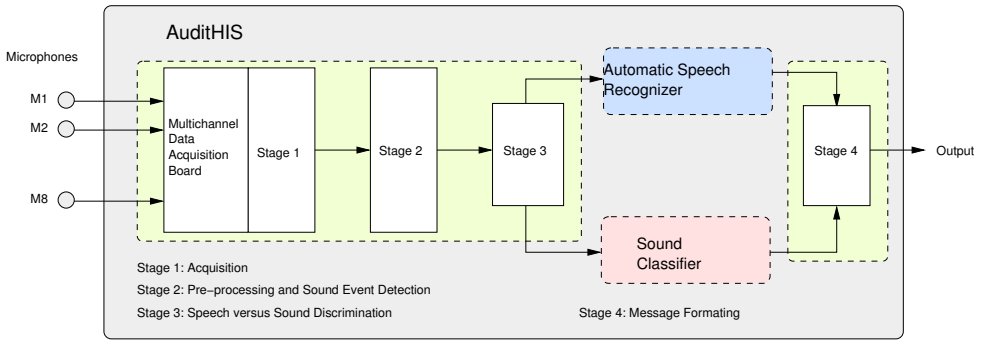

Figure 5: The AuditHIS sound analysis system.

training corpus is made of 1985 sounds and its total duration is 35 min 38 s. The sound classes are: dishes sounds, door lock, door slap, glass breaking, object falls, ringing phone, screams and step sounds. These probabilistic models can easily be extended to include more daily living sounds.

The National Instrument multichannel data acquisition board PCI-6220 is driven by the Nikal 2.1 layer (National Instrument kernel adaptation layer). The sampling rate is 16 kHz and the resolution is 16 bits.

The AUDITHISkernel is composed of 4 independent POSIX threads that exchange data through circular buffers or a queuing system.

1. **Acquisition**, in charge of transferring data from the buffer of the multichannel acquisition board to a circular buffer; this thread is dedicated solely to this task because it must take immediate actions when the card indicates that the buffer is full, otherwise the corresponding data would be lost.

2. **Pre-processing and Sound Event Detection**, that operate in parallel on each acquisition channel on the data available in the circular buffer. It is in charge of Signal to Noise Ratio (SNR) estimation on each channel and it estimates the beginning and the end of each audio event on each channel. It tries to manage the simultaneous audio events. The first step detects the portion of signal that corresponds to a sound segment. It evaluates the background noise of the room and determines a threshold of detection from this. To do this the signal goes through a wavelet decomposition (3 levels) and if the energy of the wavelet tree exceeds the adaptive threshold, the signal is recorded until its energy becomes lower than a second adaptive threshold. Each event is stored in a file for further analysis by the segmentation and recognition modules. The complete method for the detection of the bounds of a given event and also the associated evaluations is described in (Vacher et al., 2004). If the SNR is lower than a fixed value, generally +5dB, the data are not fully processed because the recognition error rate is too low for this noise level to ensure good recognition performances.

3. **Discrimination between speech and sound** is a very important task for which a classifier classifies each audio event as being speech or sound of daily living; in case of speech, data are processed by the Automatic Speech Recognizer (ASR) and in case of sound by the Sound of daily living classifier. For this reason, if the performance of this module is bad, the system will work very poorly because the recognition system (ASR) must recognize signals that are not speech or because speech signal must be analysed by the sound classifier and not by the ASR. The segmentation module is a Gaussian Mixture Model (GMM) classifier which classifies each audio event as everyday life sound or speech. The segmentation module was trained with an everyday life sound corpus (Vacher et al., 2007) and with the Normal/Distress speech corpus recorded in our laboratory (Vacher et al., 2008). Acoustical features are Linear-Frequency Cepstral Coefficients (LFCC) with 16 filter banks; the classifier uses 24 Gaussian models. These features are used because life sounds are better discriminated from speech with constant bandwidth filters, than with Mel-Frequency Cepstral Coefficients (MFCC), on a logarithmic Mel scale (Vacher et al., 2006). On the contrary, MFCC are the most widely used features for speech recognition. Acoustical features are evaluated using frames whose width is of 16 ms, with an overlap of 50%.

4. **Message formatting** is in charge of the output emission. The messages are taken into account by the Decision Stage of the Sweet-Home system: mainly voice orders but recognized sounds may be useful for context estimation. For example, the localisation of the speaker in the flat is needed to determine which lamp to light on when the speaker says “please light on”. The localisation may be estimated from sensor information and from audio analysis (Chahuara et al., 2011).

6 Automatic Speech Recognition in Smart Homes

In Sweet-Home speech recognition performed by using an ASR system. This system feeds dedicated module recognizing vocal commands or a distress situations from the decoded utterances. To address the issues of the SWEET-HOME context (noise, distant-speech) and to benefit from the multiple microphones we proposed to experiment some state-of-the-art and recent techniques that fuse the streams at three independent levels of the speech processing: acoustic signal enhancement, decoding enhance and ASRs output combination.

6.1 The Automatic Speech Recognizer

The used ASR system is the LIA (Laboratoire d’Informatique d’Avignon) speech recognition tool-kit Speeral. Speeral relies on an A^* decoder with HMM-based context-dependent acoustic models and trigram language models. HMMs are classical three-state left-right models and state tying is achieved by using decision trees. Acoustic vectors are composed

of 12 PLP (Perceptual Linear Predictive) coefficients, the energy, and the first and second order derivatives of these 13 parameters.

6.2 Acoustic models

The acoustic models were trained on about 80 hours of annotated speech. Furthermore, acoustic models were adapted for each speaker by using the Maximum Likelihood Linear Regression (MLLR) and an associated annotated corpora: MLLR adaptation is a good compromise when only a small amount of annotated data is available.

In all the presented experiments, acoustic models are firstly adapted using Maximum Likelihood Linear Regression.

6.3 Language model

For the decoding, a 3-gram language model (LM) with a 10K-word lexicon was used. The LM is computed by an interpolation of a generic LM and a specialized LM. The *generic* LM was estimated on about 1000M of words from French newspapers and broadcast news manual transcripts. The *specialized* LM was estimated from the sentences that the participants had to utter during the experiment (vocal orders, casual phrases, etc.).

6.4 Experimental context

Experiment were conducted to acquire a speech corpus composed of vocal orders utterances, distress calls and casual sentences. 21 persons (including 14 mens) participated to a 2-phase experiment to record, among other data, speech corpus in a daily living context. The average age of the participants was 38.5 ± 13 years (22-63, min-max). The participants were asked to perform several daily living activities in the smart home. No instruction was given about their behavior. Consequently, no participant emitted sentences directing their voice to a particular microphone. Sound data were recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card (Vacher et al., 2011b).

The first phase (**Phase 1**) consisted in following a scenario of activities without condition on the time spent and the manner of achieving them (having a breakfast, simulate a shower, get some sleep, clean up the flat using the vacuum, etc.). During this first phase, participants uttered forty predefined casual sentences on the phone (e.g., “Allo”, “J’ai eu du mal à dormir”) but were also free to utter any sentence they wanted (some did speak to themselves aloud).

The second phase (**Phase 2**) consisted in reading aloud a list of 44 sentences whose 9 were distress sentences and 3 were vocal orders. This list was read in 3 rooms (study, bedroom, and kitchen) under three conditions: with the vacuum on, with the radio on (vacuum off) and without noise (everything off). Only the clean condition will be used in

this paper, the noisy condition records were acquired for other kinds of experiment.

Finally, the Sweet-Home speech corpus is made of 862 sentences (38 minutes 46s per channel in total) for **Phase 1**, and 917 sentences (40 minutes 27s per channel in total) for **Phase 2** all from 21 speakers. The average SNR (Signal-to-Noise Ratio) for the considered sentences is 20.3 dB. Each sentence has been annotated.

The **Phase 1** corpus was used for development and training whereas the **Phase 2** corpus served for the evaluation.

6.5 Experiments at acoustic level

The acoustic signal enhancement can be performed by using a weight and sum algorithm: a simple sum of signals would result in a worse single channel with echoes. Weight and sum algorithm involves low computational cost and combines efficiently acoustic streams to build an enhanced acoustic signal. Given M microphones, the signal output $y[t]$ is computed by (1).

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

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

6.6 Experiments at decoding level

At the decoding level, we applied the Driven Decoding Algorithm (DDA). DDA aims to align and correct auxiliary transcripts by using a speech recognition engine (Lecouteux et al., 2006, 2008). This algorithm improves system performance by taking advantage of the availability of the auxiliary transcripts.

DDA acts at each new generated assumption of the ASR system. In order to use auxiliary transcript information, the linguistic part of the primary ASR cost function is reevaluated according to a transcript-to-hypothesis matching score α. This mechanism drives the search by dynamically rescoring the language model value, according to the alignment and word confidence scores.

The matching score denoted α is based on the number of words ϵ in the language model short-term history of size δ that are correctly aligned with the auxiliary transcript. α is

greater when the trigram is aligned and linearly decreases with the misalignments of the history:

$$\alpha = \frac{\epsilon}{\delta} \tag{2}$$

where δ is the size of the history used to compute the matching score (3 in our case)

Finally, the rescoring score α is used to fudge the linguistic probability:

$$\tilde{P}(w_i|w_{i-1}, w_{i-2}) = P^{1-\alpha}(w_i|w_{i-1}, w_{i-2}) \tag{3}$$

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

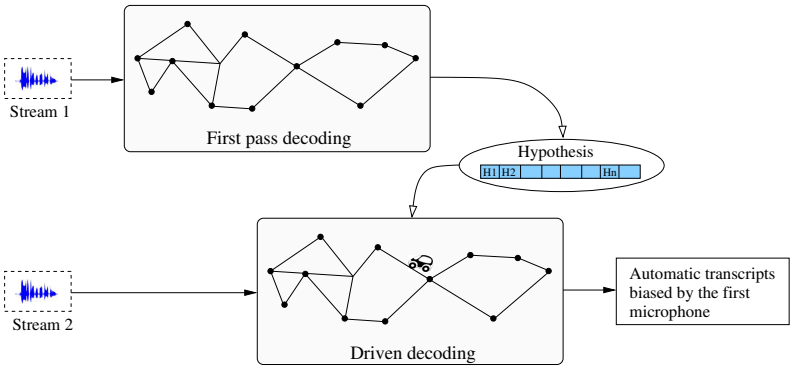

Figure 6: **DDA** used with two streams: the first stream allows one to drive the second stream

6.7 Experiments at decoding level using a priori knowledges

We propose to use the Driven Decoding Algorithm where the output of the first microphone is used to drive the output of the second one (cf. Figure 6). This approach presents two benefits:

- DDA merges the information from the two streams while voting strategies (such as ROVER) do not merge ASR systems outputs.
- The second ASR system speed is boosted by the approximated transcript (only 0.1xRT),

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

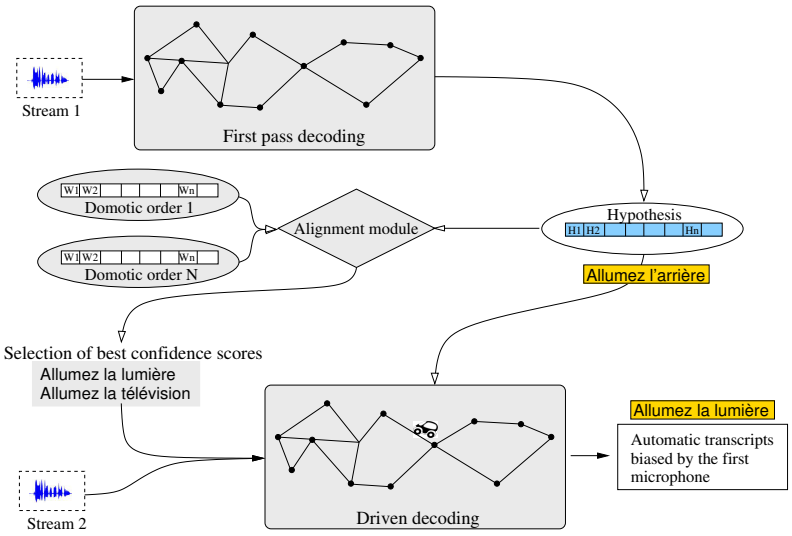

Figure 7: **DDA 2-level**: vocal orders are recognized from the first decoded stream which are then used to drive the decoding of the second stream

This approach was extended to take into account *a priori* knowledge about the expected utterances. The ASR system is driven by vocal orders recognized during the first pass. This method is called **DDA 2-level**: speech segments of the first pass are projected into the $3-best$ vocal orders by using an edit distance (cf. 5) and injected via DDA into the ASR system for the fast second pass as presented in Figure 7.

6.8 Experiments at the output level

At the ASR combination level, a ROVER (Fiscus, 1997) was applied. ROVER is expected to improve the recognition results by providing the best agreement between the most reliable sources. It combines systems output into a single word transition network. Then, each branching point is evaluated with a vote scheme. The word with the best score is selected (number of votes weighted by confidence measures). This approach necessitates high computational resources when several sources need to be combined and real time is needed (in our case, 7 ASR systems must operate concurrently).

A baseline ROVER was tested using all available channels without *a priori* knowledge. In a second time, an *a priori* confidence measure based on the SNR was used: for each decoded segment s_i from the i^{th} ASR system, the associated confidence score $\phi(s_i)$ was computed by $\phi(s_i) = 2^{R(s_i)}/\sum_{j=1}^{7} 2^{R(s_j)}$ where $R()$ is the function computing the SNR of a segment and s_i is the segment generated by the i^{th} ASR system. For each annotated sentence a silence period I_{sil} at the beginning and the end is taken around the speech signal period I_{speech}. The SNR is thus computed by (4):

$$SNR(S) = 10 * log(\frac{\sum_{n \in I_{speech}} S[n]^2}{|I_{speech}|} / \frac{\sum_{n \in I_{sil}} S[n]^2}{|I_{sil}|}) \quad (4)$$

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

6.9 Vocal orders detection

Vocal orders or distress sentences detection is performed by transcribing each vocal order and distress sentences. That gives a phoneme graph in which each path corresponds to a variant of pronunciation. Then, the number of sentences to detect in our experiments is 12 (3 vocal orders + 9 distress sentences). Automatic transcripts are transcribed in the same way.

In order to locate vocal orders into automatic transcripts T of size m, each sentence of size n from vocal orders H are aligned with T by using an edit distance algorithm at the phonetic level. The deletion, insertion and substitution costs were computed empirically. The cumulative distance $\gamma(i, j)$ between H_j and T_i is computed by (5):

$$\gamma(i, j) = d(T_i, H_j) + min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (5)$$

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

6.10 Results

The ASR stage was evaluated using the Word Error Rate (WER) and the vocal order recognition (classification) stage is evaluated using recall/precision/F-measure triplet: the number of vocal orders is about 10. Vocal orders have been manually specified by annotating all sentences. During the detection, if a marked vocal order is well detected, it is considered as detected. In all other cases we consider a detected order as a false detection. For each approach, the presented results are the average over the 21 speakers. For the sake of comparison, results of a baseline and an oracle baseline systems are provided. The baseline system outputs are composed of the best output from 7 ASR systems according to the highest SNR. The presented oracle baseline is computed by selecting the best WER for each speaker at segment level. Results are presented in table 1.

While the baseline system achieves a 18.3 % WER, all proposed approaches using SNR are able to take advantage of the multiple available microphones. The beamformin algorithm

Method	WER	Order recall	Order precision	F-measure
Baseline	18.32	88.0	90.5	89.2
Oracle Baseline	17.65	88.5	91.3	89.87
Beam Forming	16.84	89.0	92.6	90.76
DDA	12.35	92.6	97.3	94.89
DDA+SNR	11.39	93.3	97.3	95.25
DDA 2 lev.	8.93	95.6	98.1	96.83
DDA 2 lev.+SNR	8.80	95.6	98.1	96.83
ROVER	20.59	85.0	90.0	87.42
ROVER+SNR	12.20	92.7	97.4	94.99
ROVER Oracle	7.77	99.4	98.9	99.14

Table 1: WER, Vocal orders detection

allows a 8.1 % relative improvement in WER. This aspect shows that pooling all channels allows to increase the ASR system robustness. The DDA method, shows a 37.8% relative improvement by using the SNR. The 2 level DDA presents a 52 % relative improvement: this gain is easily explained by the availability of *a priori* vocal orders in the decoding pass. Finally, the ROVER weighted by SNR allows a 33.4% relative improvement but the computational cost is high (7 ASR systems).

Without using the SNR information, results present a degradation: the basic ROVER is worst than the baseline while the DDA methods are far less affected.

The baseline recognition gave a 89% F-measure. ROVER and the two DDA configurations led to a significant improvement over the baseline of about 7% absolute. Beamforming gain is not significant. ROVER performs detections not better than the DDA approaches, but requires to decode all channels. Finally, the best configuration is based on the 2 level DDA allowing a 96.83% F-measure.

7 Conclusion

Audio processing (sound and speech) has great potentials for health assessment and assistance in smart homes. Based on microphones (e.g., omnidirectional ones), it has many properties that fit the ubiquitous computing vision such as being physically intangible and freeing the user to be physically at a particular place in order to operate. Moreover, it make it possible to interact with the environment using natural language so that the user does not have to learn complex computing procedures or jargon. As such, audio technologies are particularly relevant to improve comfort via voice command and security via audio distress situations detection. However, many challenges in this domain need to be tackled to make audio processing usable and deployable in assisted living applications.

We particularly focused on two challenges in this chapter.

The first of this challenges is to develop technologies that are acceptable by the targeted users and that are useful to them. Our user study and our experience showed that microphones and audio technologies seem to be far more accepted than other more intrusive sensors. Users are particularly interested in the security enhancement this technology could provide. Interest in comfort enhancement was less clear. We did not interviewed any blind users, but it is likely that such application will be well received by this population. Despite some user studies assessing the audio interest for AAL reported in the literature, there is still many technological (Vacher et al., 2011b), usage and ethical aspects (Sharkey and Sharkey, 2012) to investigate.

The second of this challenges is to address the difficulties to develop audio-based home automation systems for the uncontrolled acoustic environment of Sweet-Home. We proposed a global architecture based on more than 10 years of experience. However, even if the audio processing issues become more clear nowadays, it is still difficult to consider all the interaction aspects that will fit the users. This is why we adopted a user centric system development as it seems to be the most sensible way of answering explicit and implicit needs.

Regarding the audio processing conditions, many interferences appear in daily usage and, in our experiments, most of the encountered problems were indeed due to noise, environmental perturbations or overlapping sound events. To tackle these problems, we proposed a method to benefit both from the multi-source environment and the *a priori* knowledge about the voice command task. The multisource data made it possible to employ several Automatic Speech Recognition (ASR) systems to reach a consensus about the actual uttered voice order. This improved dramatically the performance. However, this is when some knowledge is introduced directly within the ASR that best performances are reached. This original method, called the Driven Decoding Algorithm (DDA), has a high potential for highly constrained tasks such as voice command in a domotic context. We expect even better performance if input signals are further enhanced at the acoustic level.

Regarding this signal enhancement, we plan to conduct several studies to determine what source separation methods, such as Independent Component Analysis, can improve the recognition. Moreover, regarding speech recognition, probabilistic models need to be adapted to the ageing population. We are currently recording seniors' voice to adapt our ASR models to this population.

Acknowledgments This work is supported by the Agence Nationale de la Recherche (ANR-09-VERS-011).

References

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

Benabderamane, Y., Selouani, S., and O'Saughnessy, D. (2011). Blind speech separation in multiple environments using a frequency oriented pca method for convolutive mixtures. In *Proceedings of Interspeech 2011, Florence, Italy*, pages 557–560.

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

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, Space, Time and Ambient Intelligence*, pages 17–22, Barcelona, Spain.

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

CHiME (2011). Machine listening in multisource environments, CHiME Workshop, Interspeech2011. `http://spandh.dcs.shef.ac.uk/projects/chime/`.

Cornet, G., Franco, A., Rialle, V., and Rumeau, P. (2007). volume 703 of *Société Française de Technologie pour l'Autonomie et de Gérontechnologie*, chapter Les gérontechnologies au cœur de l'inovation hospitalière et médico-social, pages 53–58.

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

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.

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

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

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

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

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

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

Geiger, J., Lakhal, M., Schuller, B., and Rigoll, G. (2011). Learning new acoustic events in an hmm-based system using map adaptation. In *Proceedings of Interspeech 2011, Florence, Italy*, pages 293–296.

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

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.

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

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

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

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

Jambon, F., Mandran, N., Meillon, B., and Perrot, C. (2010). Évaluation des systêmes mobiles et ubiquitaires : proposition de méthodologie et retours d'expérience. *Journal d'Interaction Personne-Système*, 1(1):1–34. [in French].

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

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

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

Leng, Y., Tran, H. D., Kitaoka, N., and Li, H. (2011). Alternative frequency scale cepstral coefficient for robust sound event recognition. In *Proceedings of Interspeech 2011, Florence, Italy*, pages 297–300.

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

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

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.

N. Tran, W. Cowley, A. P. (2011). Adaptive blocking beamformer for speech separation. In *Proceedings of Interspeech 2011, Florence, Italy*, pages 577–580.

Nocera, P., Linares, G., and Massonié, D. (2002). Principes et performances du décodeur parole continue speeral. In *XXIVées journées d'étude sur la parole*. Laboratoire Informatique d'Avignon.

Noury, N., Rumeau, P., Bourke, A., O'Laighin, G., and Lundy, J. (2008). A proposal for the classification and evaluation of fall detectors. *IRBM*, 29(6):340–349.

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., Fleury, A., Vacher, M., and Noury, N. (2009). Determining useful sensors for automatic recognition of activities of daily living in health smart home. In *in Intelligent Data International Workshop on Analysis in Medicine and Pharmacology (IDAMAP2009)*, pages 63–64, Verona, Italy.

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

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

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

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.

Rodin, J. (1986). Aging and health: effects of the sense of control. *Science*, 233(4770):1271–1276.

Sarmiento, A., Durán, I., Cruces, S., and Aguilera, P. (2011). Generalized method for solving the permutation problem in frequency-domain blind source separation of convolved speech signals. In *Proceedings of Interspeech 2011, Florence, Italy*, pages 565–568.

Seymore, K., Stanley, C., Doh, S., Eskenazi, M., Gouvea, E., Raj, B., Ravishankar, M., Rosenfeld, R., Siegler, M., Stern, R., and Thayer, E. (1998). The 1997 cmu sphinx-3 english broadcast news transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA.

Sharkey, A. and Sharkey, N. (2012). Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40.

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

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

Vacher, M., Istrate, D., Portet, F., Joubert, T., Chevalier, T., Smidtas, S., Meillon, B., Lecouteux, B., Sehili, M., Chahuara, P., and Meniard, S. (2011a). The SWEET-HOME Project: Audio Technology in Smart Homes to improve Well-being and Reliance. In *33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2011)*, Boston, USA.

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

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

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

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

van Hoof, J., Kort, H., Markopoulos, P., and Soede, M. (2007). Ambient intelligence, ethics and privacy. *Gerontechnology*, 6(3):155–163.

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

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

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

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

Ziefle, M. and Wilkowska, W. (2010). Technology acceptability for medical assistance. In *Pervasive Health*.