A french corpus of audio and multimodal interactions in a health smart home

Anthony Fleury, Michel Vacher, François Portet, Pedro Chahuara, Norbert Noury

Journal on Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

A French Corpus of Audio and Multimodal Interactions in a Health Smart Home

Anthony Fleury* · **Michel Vacher** · **François Portet** · **Pedro Chahuara** · **Norbert Noury**

Received: Feb. 29, 2012, Revised: June 11, 2012, Accepted: July 3, 2012

Abstract Health Smart Homes are nowadays a very explored research area due to the needs for automation and telemedicine to support people in loss of autonomy and also due to the evolution of the technology that led in cheap and efficient sensors. However, collecting data in this area is still very challenging. As a consequence, many studies can not be validated on real data. In this paper, we present two realistic datasets acquired in a fully equipped Health Smart Home. The first is related to distress detection from speech (450 recorded sentences) and involved 10 participants, the second involved 15 participants who were performing several instances of seven activities of daily living (16 hours of multimodal data).

Keywords Health Smart Home · Environmental Sensors · Audio and Speech Analysis · Activity monitoring · Activity Recognition

1 Introduction

Recent developments in Home Automation, Robotics and ICT have led to the concept of *smart home*. A smart home is a home environment fitted with ICT technologies to enhance human machine interaction and empower the dweller. One of the most promising applications of smart home is in the health domain. Due to the growing number of elderly people which, according to the World Health Organization (WHO), is going to reach 2 billion by 2050, Health Smart Homes [5,33] were designed to improve daily living conditions and independence for the population in loss of autonomy. Indeed, Smart homes seems very promising 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. If this aim is reached, smart homes will facilitate the daily life and the access to the whole home automation system. Many studies have been conducted in different countries to define what are elderly wishes concerning a smart home system able to help them in their daily life [23,28,9,20,3,44]. However, it is still unclear which particular interfaces would be the most adapted to this population. Moreover, these people are often the less capable of using the complex interfaces due to their disabilities (e.g., cognitive decline) or their lack of ICT understanding.

Audio-based technology has a great potential to become one of the major interaction modalities in smart home and more generally in *'Ubiquitous Computing'*. As introduced by Weiser [42], ubiquitous computing refers to the computing technology which disappears into the background, which becomes so seamlessly integrated into our environment that we do use it naturally without noticing it. Audio technology has not only reached a stage of maturity (e.g., automatic speech recognition is a feature of many computers and mobile applications) but has also many properties that fit this vision. It is physically intangible and depending on the number and type of the sensors (omnidirectional microphones) that are used, it does not force the user to be physically at a particular place in order to operate. Moreover, it can provide interaction using natural language so that the user does not have to learn complex computing pro-

Anthony Fleury
Univ. Lille Nord de France, F-59000 Lille, France and
Mines Douai, IA, F-59500 Douai, France
Tel: +33 3 27 71 23 81, Fax: +33 3 27 71 29 17
E-mail: Anthony.Fleury@mines-douai.fr

Michel Vacher · Francois Portet · Pedro Chahuara
Laboratoire d'Informatique de Grenoble,
UMR 5217, CNRS/UJF/Grenoble-INP, F-38041 Grenoble, France
E-mail: firstname.lastname@imag.fr

Norbert Noury
Univ. Lyon, lab. INL, UMR CNRS/UCBL/ECL/INSA 5270
F-69621 Villeurbanne, France
E-mail: Norbert.Noury@insa-lyon.fr

cedures or jargon. It can also capture sounds of everyday life which makes it even easier to use (hand clapping to control light is a well known example) and can be used to communicate with the user using synthetic or pre-recorded voice. Despite all this, a relatively small number of smart home projects have seriously considered audio technology and notably speech recognition in their design [15,2,16,27,11,26]. Part of this can be attributed to the fact that this technology, though mature is still complex to set up in a real environment and to the fact that important challenges still need to be overcome [40].

Research studies which concern smart home ask for a large scope of skills and a high amount of resources. One of the main problems that impede research in this area is the need for a big amount of annotated data (for analysis, machine learning and reference for comparison) while so few are available with the correct experimental settings. The acquisition of such datasets is highly expensive both in terms of material and of human resources. For instance, in the SWEET-HOME project [35] , the acquisition and the annotation of a 33-hour corpus involving 21 participants in a smart home cost approximatively 70k€ to the society. Thus, making these datasets available to the research community is highly desirable. In this paper we describe a multimodal corpus acquired in a audio-based smart home that we have made available to the community [17].

The aim of this paper is to clearly present the work that has been done to create the two datasets that we are making available on-line, and to discuss about some possible uses of these datasets for other researchers.

The paper is organized as follow. First, Section 2 starts with a description of the concept of smart homes, with the issues that are addressed. Then, Section 3 develops a review of the different datasets in the field of smart homes that can be available on-line to researchers. After this section, the description of our datasets starts with the technical details of the environment in which it has been acquired (Section 4), followed by the acquisition settings and the corpus recording for two different experiments (Section 5). We acquired and we now disseminate two different corpora including respectively 10 and 15 participants, that are then presented in Section 6. Finally, Section 7 deals with the annotation scheme that has been used for these corpora. To illustrate how they have been used, different experiments are summarised in Section 8. Finally, this paper finishes with a discussion on the interest of the corpus for the community (Section 9) followed by a short conclusion (Section 10).

2 The Health Smart Home Context

Health Smart Homes are homes fitted with sensors and actuators to monitor the behaviour of the dweller and her interactions with the environment in order to act on the environment to provide the adequate assistance or adjustment. In the following, we present how audio interfaces in smart home can enhance security and contribute to provide context awarness.

2.1 Security Enhancement Through Audio Interfaces in Smart Home

Security enhancement in-home is a prevalent concern for the elderly population and their relatives [18,20]. This enhancement has been typically tackled with worn-sensors or push-button appliance with limited success until now. The main problems of these systems is the fact that the person can forget to wear them or to change the batteries. Indeed, this necessitates that the person is constantly wearing it or is constantly aware that she must wear the sensor (e.g., every morning). However, it is clear that concerning persons with cognitive decline or changing mood this is inappropriate. Regarding personal emergency response systems, studies showed that push-button based control is not adapted to this population [16]. Indeed, a tactile system would necessitate being physically available at the location where the command has to be activated, or would imply for the person to constantly know where the remote controller is. There is thus a strong interest in voice interfaces for the elderly. Indeed, it does not necessitate to be at a particular place to operate, does not necessitate to wear it and is especially adapted to the physically disabled persons (e.g., person in wheel-chair or blind). Moreover, audio interfaces make speech interaction possible which is the most natural way of interacting. This has a direct impact on security enhancement by avoiding dangerous situation. 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 and thus avoid a potential fall.

In this paper, we concentrate on distress situations. These concern all situations in which the person is (or feel to be) in danger and needs rapid assistance to get out of this situation. It is the case, for instance, when the person falls, when the person has a panic attack, when the person is feeling faint.... In these case of distress situations, audio interfaces permit to capture highly semantic information such as speech or gasps. For instance, after a fall or an attack, if the person remains conscious but can not move, the system offers the opportunity to call for help by the voice. Moreover, during a fall or an attack, objects can falls, gaps can be heard and many others sounds that can inform about the situation and the state of mind of the person.

Setting up a system that will detect all of these various distress situations is very challenging, this is why we focused on the automatic recognition of vocal *calls for help*. When a person is in distress and when a vocal system is present she has the opportunity to utter distress sentences

(e.g., *help, call a doctor* . . .). Two information channels must be considered in this case: the meaning of the utterance and the emotion conveyed with it. In general, it could be very profitable to evaluate the emotion level in speech in case the speaker feels a great emotion, the system could take this information into account before decision making. For instance *"call my daughter"* can be a neutral or urgent command depending on the way the person uttered it. As pointed out by gerontologists [7,36] and physicians, taking the emotional aspect into account is important not only for health assessment but also to move towards *emphatic* houses. It appears that emotion detection is as important as the recognition of the sentence itself because this is a good indication of the state of mind of the person. However, emotion recognition from speech is still highly challenging. Moreover, in the audio signals, it must be distinguished what is genuinely produced by the speaker from a various set of interferences that may corrupt the signal being analysed. In order to build the technology for this kind of audio interfaces a large amount of data sets must be acquired that are sufficiently diverse in term of population, premises, language, culture, etc. to develop and test new processing techniques. As stressed in the introduction few project considered speech interface in their setting and even less consideration was given to distress call. As a result, we are not aware of any freely available dataset usable for this aim. In this paper, we introduce the experiment that we conducted in order to acquire such dataset and to make it available to the research community.

2.2 Activity Recognition for Context Aware Interactions in Smart Home

In the Ambiant Intelligence domain a whole field of research is concerned with making the house more and more intelligent by going from telecare application to proactive houses that can learn from experience and understand the user's intention and state of mind. To do so, the house must recognise or understand the situation in which the user is. This kind of decision is related to the *context-aware computing* domain [29]. There are many definitions of what context means or should mean but it seems that context is most wildly accepted as being the conditions in which an action is defined and executed [10]. Context may contain a various set of information ranging from those explicitly provided by the user (e.g., age, profile, native language. . .) to those which are implicit and must be extracted from the sensors (e.g., state of mind, activity, position. . .). To illustrate this *context-aware* interaction, let's consider the following very simple example. Suppose a person in her bedroom vocally ordering to the smart home "*turn on the light*". Based on the utterance the system should be able to understand the order but the information provided would be insufficient to know how to act. Indeed, there might be several lights in the room (e.g., a ceiling light and a bedside lamp) and the intensity of the light might be controllable. If the person is reading in her bed, the bedside lamp could be turned on with a medium intensity. If the person is cleaning up the room, the ceiling light at full intensity might be the most adequate action. If the person is waking up in the middle of the night the ceiling light at low intensity is the most relevant action, and so on. It is clear that if the context is wrongly estimated (or not considered) this can result in dazzling the waking up person or in any other inappropriate actions. As illustrated by this example, proper assessment of the activity is central to set up a system reactive to vocal order. In this why, the following concentrates on this task.

ADL	iADL
Feeding	Using the phone
Dressing	Handling money
Going to the toilets	Doing laundry
Continence	Preparing and managing food
Hygiene	Handling medication
Locomotion	Doing shopping
	Using public/private transports
	Maintain accommodation

Table 1 Example of Activities of Daily Living (ADL) and Instrumental ADLs

Apart from context provision, activities can also be used to monitor the person's behaviour. Indeed, activities are the atomic elements of a person's day and automatic recognition of them could permit to identify life rhythm and deviances from his/her usual daily routine. Many taxonomies of activities exist in the literature. However, in our research, we were primarily interested in the important activities of the elderly population. Thus, the activities were defined with respect to the tools that geriatricians use. A traditional tool is a questionnaire based on the index of Activities of Daily Living (ADL) [22] and on Instrumental Activities of Daily Living (iADL) [24]. ADL and iADL scales are described table 1. While ADL is related to daily human needs, iADL is focused on the activities that involve handling of tools (e.g., preparing food) or higher level organisation (e.g., handling money).

As with distress detector, the development of this technology needs a very large amount of representative data. But contrary to the distress situation detection, there are several datasets available (the reader can find a short overview of these in the next section). However, very few of them contain audio information and even less have good quality recordings of non-worn microphones (i.e., no head set). This is not surprising given that most research undertaken in activity recognition is strongly (if not essentially) based on video cameras [1]. Our approach is completely different as video cameras seem to pose acceptance and ethical problem when use in intimate life [35,37], though means exist to

reduce their intrusion level [30]. Thus, apart from audio interfaces, we considered classical home automation sensors such as presence infra-red detectors and contact doors. In this paper, we introduce the experiment that we conducted in order to acquire such dataset and to make it available to the research community.

3 Review of the Current Available Corpora in Smart Home

Development of smart home technologies requires datasets obtained from experimental platforms providing the necessary conditions to represent real situations. Collecting these datasets is, in the most cases, very demanding in effort, resources, and time, and ask for a very good know-how in order to get the most naturalness possible from the experiment. Given the cost of such acquisition, these resources are often collected as part of a funded project (either national research bodies or companies) which naturally biases the collection towards the project objectives (e.g., validation of a specific interface). Moreover, the experimental material are often available to a small part of the research community. As a result, this slows the research advances in this domain. To overcome these limitations a number of corpora in Smart Home collected from experimental platforms in several projects have been made available to the research community. A few websites appeared to list some of them such as BoxLab supported by the MIT [1] or from the university of Hong Kong [2]. In this section, a small number of such projects are introduced.

In the MavHome project [43] the objectives were to use data mining techniques to control the pervasive environment according to the inhabitant's activities. To this end, they acquired two datasets obtained from an office-like environment and an apartment testbed. The former corpus contains sensor events collected during two months where six students worked in the office, whereas the testbed corpus provides sensor data of a single person who lived there for a two-month period. Both corpora are composed of text files with a list of tuples including the time of the event, the source sensor, its location, and its state value. The installed sensors mainly informs on motion, light, temperature, humidity and doors state.

In the University of Amsterdam, Kasteren et al. [21] applied statistical models to perform activity recognition using a corpus collected in a normal apartment adapted to this purpose. In this case, 14 state-change sensors where installed in the apartment where a 26-years-old man lived alone during 24 days. The sensors were installed on objects that can easily be associated with ADLs such as the microwave, the fridge, and the toilet flush. Each sensor event was annotated to relate it to a specific activity. However, these annotations were made by the inhabitant himself who used a headset along with a speech recognition module. At the beginning and end of each activity the inhabitant gave a vocal command whose interpretation established the boundaries of activities in the annotated data. This could put into question the naturalness of the dataset given that the person could not forget at any stage that he was not in a normal situation.

In the CASAS project [6] of the University of Washington, several corpora were collected in a flat located in the university campus. As in the previous case, the research focused in the recognition and analysis of ADLs. Likewise, the data contains a list of sensor events chronologically sorted and information about the beginning or end of activities. One of the corpora was acquired from 20 people performing activities within some hours, then other experiments were carried out with a single person living in the apartment for several months. With regard to the employed sensors, they fall in the categories of motion, contact doors, temperature, electricity and water consumption.

The fact that the above described corpora does not include videos or audio recorded during the experiments makes them more difficult to analyse and to annotate some other aspects of inhabitant behaviour than the one they were designed for. The MIT's Place Lab smart home [19] does offer several corpora containing video and audio files of the experiments. The main corpus they provide, PLIA1, contains sensor data and annotation of 84 activities. The amount of data of this corpus is extensive since 214 sensors were installed in the experimental flat measuring temperature, light, humidity, water and gas flow. However in spite of the big size of the corpus, its main drawback is that it only covers 4 hours of experiment, which is a too short period to obtain enough information for researches on activity recognition or decision making.

In the GER'HOME[3] [45] by the INRIA and CSTB, the aim was to detect ADLs of elderly persons according to several sensors such as video cameras, boolean sensors (door contacts, pressure detector) and consumption meters (water-meter). The activity recognition was performed through an expert system [41] and the video processing is supported by a 3D model of the flat. The corpus does not include audio data but it is one of the few which contains actual elderly users.

From this very short overview it can be noticed that the audio channel in the house is under-considered. This is why we acquired the corpus described in the following and made it available to the community.

[1] boxlab.wikispaces.com/

[2] www.cse.ust.hk/~derekhh/ActivityRecognition

[3] gerhome.cstb.fr

4 Technical Environment and Sensors

The health smart home considered in this study was built in 1999 by the TIMC-IMAG laboratory in the faculty of Medicine of Grenoble [32]. This real flat of 47m^2 is composed of a bedroom, a living-room, a hall, a kitchen (with cupboards, a fridge...), a bathroom with a shower and a cabinet. An additional technical room contains all the materials required for data recording. The flat is depicted in Figure 1.

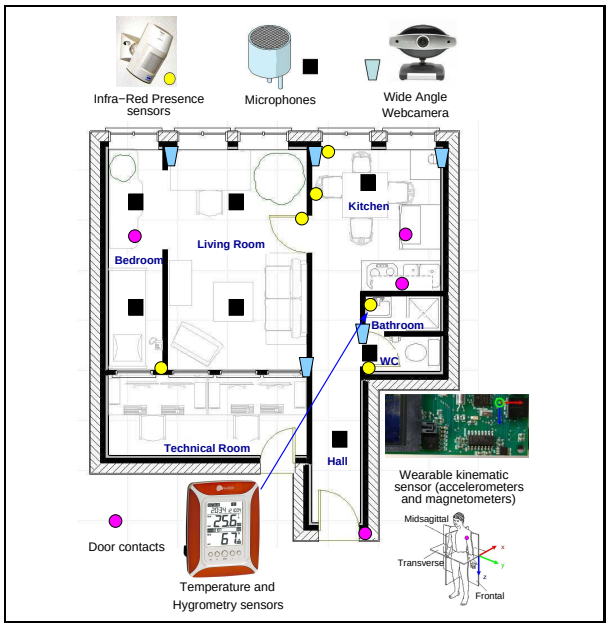

Fig. 1 Map and location of the sensors inside the Health Smart Home of the TIMC-IMAG Laboratory in the Faculty of Medicine of Grenoble.

The basis of the flat is a controller collecting information from wireless PID (Presence Infrared Detectors), wireless weight scale, oxymeter and tensiometer. The exchange protocol is the Controller Area Network (CAN) bus protocol. All the data are stored in a SQL database. Since 2000, the TIMC-IMAG laboratory has been working with the Laboratoire d'Informatique de Grenoble to add audio sensing technology in their health smart home. Omni-directional microphones were installed in the ceiling directed vertically to the floor and specific software were developed to record and analyse the audio channels in real-time. Furthermore, several webcams have been placed in the home for the purpose of marking up the person's activity and to monitor the use of some furniture (here the fridge, the cupboard and the chest of drawers). The real-time recording of these sensors is shared between four computers in the technical room:

1. The first one is devoted to the sound and speech analysis. It is an Intel Xeon 3.4 GHz based computer with 4 GB of RAM and a GNU/Linux OS. It is equipped with a National Instrument acquisition board (National Instrument PCI-6034E) to record simultaneously the seven microphone channels of the flat.
2. The second one is dedicated to the capture of three of the USB2 web-cameras and is also receiving the data from the CAN bus of the flat. This one and the next one are Intel Pentium IV 3 GHz based computers with 4 GB of RAM and a MS Windows XP OS.
3. The third one is collecting the data from the two other USB2 web-cameras and from the systems which collects the temperature and hygrometry parameters in the bathroom.
4. The last one is in charge of the door contacts of the kitchen and the bedroom. It is an Intel Centrino 1.8 GHz with 2 GB of RAM with a MS Windows XP OS.

The audio channels are processed by the AUDITHIS audio system (developed in C language) which is running in real time [38]. The general organization of the audio system is displayed Figure 2, height microphones are set in the ceiling as shown in Figure sec:techenv. AUDITHIS is parametrised through a dedicated module, while other modules run as independent threads and are synchronized by a scheduler. The 'Acquisition and First Analysis' module is in charge of data acquisition of up to 8 analogue channels simultaneously, at a sampling rate of 16 kHz. Each time the energy on a channel goes beyond an adaptive threshold, an audio event is detected by the 'Audio Detection' module. A record of each audio event is kept and stored in MS-Wave format on the hard drive of the computer. For each event, Signal-to-Noise Ratio (SNR), room, date and duration are stored in an XML file. In this way, the AUDITHIS can have three different uses:

1. real-time analysis;
2. recording of sound in a realistic environment, these data are useful for learning models more adapted to the context and that can be used for future experiments;
3. these recorded data can also be used for new analysis using other algorithms or improved models. An example is speech recognition by the Automatic Speech Recognizer (ASR) Raphael included in AUDITHIS. This recognizer is trained by large corpora like BREF120 recorded by more than 120 speakers, and the sentences recorded by AUDITHIS may be used to evaluate other ASRs or more specific models...

The PIDs (Atral DP8111) sense a change of temperature in their reception cone. They are mainly used for alarm systems and lighting control. Each movement in a determined zone generates a detection that is transmitted through the CAN bus to the technical room PC. Six PIDs have been placed inside the flat. There is one in the kitchen that monitors the space around the table and the cooking place; one in the living room to monitor the sofa; one in the bedroom

to monitor the bed; one in the bathroom; one in the toilets and one to monitor the entrance hall. Analysing the time series (and its evolution) of detections for the location sensor can give relevant information for determining the mobility (transitions between sensors) and the agitation (number of successive detections by a same sensor) of the person [25].

Three door contacts have been placed inside furniture (cupboard, fridge and convenient). They are simulated by video cameras. Each frame is thresholded to detect the status 'open' or 'closed'. The output for these sensors is the time of occurrence of the transitions.

The last home sensor delivers information on temperature and hygrometry. To be informative, this sensor is placed inside the bathroom to detect a shower. During this activity, the temperature will rise (as the person takes a hot shower), and the hygrometry will also increase. The sensor used is a commercial product (La Crosse Technology, WS810), it measures both information every five minutes.

Together with the fixed home sensors, a wearable sensor is available. ACTIM6D, our home made kinematic sensor, is a circuit board equipped with a three axis accelerometer (MMA7260Q, Freescale) and a three axis magnetometer (HMC1053, Honeywell) [12]. It is kept tight on the subject and creates a new referential in which the movements of the person can be analysed. The position and the orientation of this kinematic sensor is shown on Figure 1. This sensor outputs text files containing the timestamps of the changes of posture and the beginning/end times of each recognized walking sequences.

Finally, video recording was added for two purposes. The first is to create an index of the different activities performed and the second is to simulate new sensors or create a gold standard for one of the sensor in the dataset.

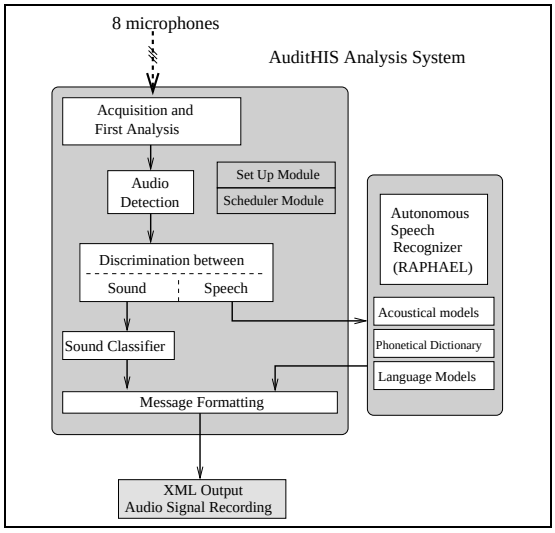

Fig. 2 The AuditHIS and RAPHAEL systems

5 Corpus Collection

This section presents the settings of the two experiments that were conducted in the health smart home to acquire audio and multimodal interactions in a domestic environment. The first one is related to distress call by the dweller in emergency case, for this, only microphones were used. The second one is especially related to everyday life of the dweller and all sensors were taken into account. Seven activities were considered in this study: (1) Sleeping; (2) Resting: watching TV, listening to the radio, reading a magazine…; (3) Dressing and undressing; (4) Feeding: realizing and having a meal; (5) Eliminating: going to the toilets; (6) Hygiene activity: washing hands, teeth …; and (7) Communicating: using the phone.

5.1 Distress Call Corpus Acquisition

This experiment was designed to capture speech utterances when the dweller is in a distress situation (simulated) and is calling for help. For evaluation purpose, neutral colloquial utterances and voice command expressions were also included in the scenarios. Therefore, 3 categories of sentences (in French) were considered: home automation orders, distress calls and usual phone conversations. To precise the vocabulary, some of the experiments use only the differentiation between distress and non-distress events, in that case both usual phone conversation and home automation orders are in the same class of "normal" sounds opposed to the distress call sentences. In this experiment, every participant executed a scenario composed of 45 sentences to utter (20 distress sentences, 10 normal sentences and 3 phone conversations of 5 sentences each). Excerpts of the list of sentences are provided Table 2. This corpus has been acquired in conditions that are as closed as possible to reality (due to the difficulty to really acquire some distress situations in real life). Considering its size, it is not designed to train some classifier but it is designed to test existing language models in extremely difficult situations (without embedded microphones and in uncontrolled conditions) in order to evaluate the detection of distress keyword in daily living environment.

Ten healthy volunteers (including 3 women) were recruited for the experiment (mean age: $37.2 \pm SD = 14$ *years*,

Domotic Order	Distress Sentence	Usual Phone Conversation
Allume la lumière	*À l'aide*	*Allô c'est moi*
Éteins la lumière	*Je suis tombé*	*Allô c'est qui*
Ferme la porte	*Une infirmière vite*	*Bonjour Monsieur*
Ouvre la porte	*Appelez une ambulance*	*Dehors il pleut*

Table 2 Some examples of French sentences for the 3 considered categories

mean weight: $69 \pm SD = 12$ *kgs*, mean height: $1.72 \pm SD = 0.08$ *m*). The protocol was quite simple. Each participant was alone in the flat. The participant had to go to the living room and to close the door. Then, s/he had to move to the bedroom and read aloud the first half of one of the five successions of sentences, out of 10 normal and 20 distress sentences. Afterwards, the second half of the set of sentences had to be uttered in the living room. Finally, each participant was called 3 times and had to answer the phone and read the predefined randomly selected phone conversations (5 sentences each).

The sentences were uttered in the flat, with the participant sat down or stood up. To validate the system in uncontrolled conditions, they were free to place themselves away from the microphones (in practice it was between 1 and 10 meters) and they had no instructions concerning their orientation with respect to the microphones (they could choose to turn their back to the microphone direction). Moreover, the experiment took place during daytime – hence we did not control the environmental conditions of the experimental session (such as noises occurring in the hall). The microphones were set on the ceiling and directed vertically to the floor as shown on Fig. 1. A phone was placed on a table in the living room but it was not recorded, only the microphones of the ceiling were used.

5.2 Everyday Audio and Multimodal Interactions Corpus Acquisition

This experiment was designed to acquire data of natural interactions between a dweller and her/his natural domestic environment, these interactions being captured by the home and wearable sensors described in Section 4. The original aim of the study was to use the traces of the interactions in the sensor data to automatically recognize activities of daily living (see Section 2.2 and [14]) so that human computer interface can be used in a context aware way. Seven activities were considered in this study: (1) Sleeping; (2) Resting: watching TV, listening to the radio, reading a magazine…; (3) Dressing and undressing; (4) Feeding: realizing and having a meal; (5) Eliminating: going to the toilets; (6) Hygiene activity: washing hands, teeth …; and (7) Communicating: using the phone. Fifteen healthy participants (including 6 women) were asked to perform 7 activities, at least once, without condition on the time spent. Four participants were not native French speakers. The average age was 32 ± 9 years (24-43, min-max) and the experiment lasted from 23 minutes 11s to 1h 35 minutes 44s. Figure 3 shows a translated excerpt of the information given to the participants. A visit, before the experiment, ensured that the participants would find all the items necessary to the seven ADLs. Participants were free to choose the order of execution of the ADLs to avoid repetitive patterns. Participant were also asked not to act any characters. Data were marked-up afterwards using the acquired video.

```
You enter the apartment and, from that moment, you
will be alone. You will leave the flat at the end of
the experimental session. You will have to perform the
following tasks of daily living in whatever order you
wish:

- Sleeping: you lie in the bed and stay ‘‘asleep’’ for
  a reasonable time (of your choice)
- Resting: you sit on the couch or a chair in the
  living room and perform the activity that you
  like to feel relaxed (watching TV, reading, doing
  nothing...)
- Feeding: you prepare a breakfast with the equipment
  and the ingredients in the kitchen cupboard and then
  eat (using the kitchen table). Then, you wash the
  dishes in the kitchen sink.
- Hygiene: you wash your hands, face, and pretend to
  brush your teeth in the bathroom.
- Toilets: you pretend to go to the bathroom (sit on
  the toilet, flush the toilet...).
- Dressing: you put and remove the clothes in the
  chest of drawers near the bed (over your own
  clothes). You can do this activity before going
  to sleep and after sleep or after performing the
  hygiene task.
- Communication: you will be called over the phone (in
  the living room) several times. Each time you will
  read the sentences indicated near the phone.

If possible, you are asked to perform each activity
at least once and during at least three consecutive
minutes.
```

Fig. 3 Translation of an excerpt of the explanation sheet that participants had to read and to keep with them during the experiment

Apart from the fact that participants were not in their own home, a few other factors such as the video cameras might have influenced their behaviour. Before signing the agreement form, participants were shown what was acquired by the different cameras. One point that was particularly discussed with the participants was the camera recording the door of the toilets. This camera was clearly aiming the top of the door and the ceiling to respect participant's privacy. A code was established with them: when the person was in the bathroom/toilets area if this door was partially closed, then it indicated an elimination activity, otherwise s/he was performing the hygiene activity. This is the less intrusive way to indicate the intimate activities being performed we found. Despite these constraints, overall, the participants were quite relaxed and did not show any hesitation in making themselves comfortable in the flat by actions we had not anticipated (e.g., opening the window, adjusting the blind, etc.). This has led to unexpected situations with, for instance, a participant that had a phone call during the experiment and answered it naturally. Moreover, the recording was not cancelled or moved because of external conditions. For instance, the recording of a participant was performed

during a thunderstorm which increased the number of detected sounds. That is why, despite the fact that the experiment could have been conducted in a more controlled way, we believe that this piece of freedom has led to a natural corpus which would be hard to reproduce.

Regarding the sensors, this flat represented a very hostile environment similar to the one that can be encountered in real home. This is particularly true for the audio information. The sound and speech recognition system presented in [38] was tested in laboratory and gave an average Signal to Noise Ratio of 27dB. In the Health Smart Home, the SNR average is lower and reaches 12 dB. Moreover, we had no control on the sounds that are measured from the exterior of the flat, and a lot of reverberation was introduced by the 2 important glazed areas opposite to each other in the living room. More details about the experiment can be found in [14].

Speaker	SNR Living-room (dB)	SNR Bedroom (dB)
1	18.5 (38)	15.1 (7)
2	16.4 (36)	14.7 (9)
3	21.1 (38)	18.4 (7)
4	17.9 (36)	16.6 (9)
5	17.4 (31)	14.9 (12)
6	15.7 (28)	14.1 (15)
7	12.5 (32)	10.7 (8)
8	11.7 (35)	13.7 (6)
9	11.0 (40)	12.0 (5)
10	12.9 (35)	11.1 (9)
Average	15.5 (34.5)	14.2 (8.7)

Table 3 Mean SNR in dB for the speech corpus composed of distress and normal sentences uttered by 10 different speakers. The number between parenthesis is the number of sentences taken from each microphones.

6 Acquired Corpus

From these two experiments, two sets of data, described in this section, were acquired and are freely available [17].

6.1 Distress Call Corpus

The distress call corpus is composed of data from 10 different speakers acquired following the method described in Section 5.1. In this way, 3.546 sounds occurrences containing the 45 sentences to be uttered by the participants were recorded but also perturbing sounds generated from within or outside the flat (e.g., the engine of the elevator next door). This amount of audio events was filtered in order to keep only the sound that corresponded to the participant speech. As the recording was multisource, only the sound event with the highest SNR was considered as the best utterance.

Finally, the corpus contains 450 uttered sentences, each in a separate wave file, with the corresponding transcription files. The acquisitions were made by the software on-the-fly, with a specific algorithm, described in our previous papers, for the detection of events (this software is designed to allow on-line distress situation detection). As a consequence, a complete wave file of the session is not available, only the different detected sounds can be downloaded.

For the experimentation, the participant was in the living-room. As a consequence, the sounds that had the best SNR were taken from the microphones of the living-room (close to the window), and the others were from the one of the bedroom (close to the technical room). Table 3 shows the average value of the SNR of the sounds acquired for each of the participant and each of the two microphones. For the items in which we do not have exactly 45 sounds, it means that the other ones have been either never classified as speech by the software or not even recorded (too noisy environment).

6.2 Multimodal Daily Living Corpus

This corpus has been acquired during summer 2008, and contains about 16 hours of data from 15 different participants. These sensor data (without the video recordings) represents a total of 3.53 GB (in more than 44,000 files). The video by itself represents 8.61 GB of data (with one video that contains the four camera for each person). In the following the data acquired from each kind of sensor is detailed.

6.2.1 PID

More than 1700 firings have been recorded. Surprisingly, the sensitivity of the PID was not as good as expected. The sensitivity to detect a change of rooms in 10 records is 80%. To succinctly recall the functioning of PIDs, they detect perturbations of a background temperature (estimated via infrared radiations) thus a person walking between the PID sensor and an area at a different temperature (such as a wall or the floor) will trigger the sensor. The problem of missing detections could be explained by the fact that the experiments have been done in summer. Thus, the difference between the temperature of the wall in the flat and the one of the participants' body would not have been sufficient to provoke an infra-red energy change in the sensor. This would be especially true when the movement is rapid. However, this problem reflects that no source is 100% reliable and that PIDs, though largely used in smart home, should be supplemented by other location related sensors (e.g., audio).

6.2.2 Doors Contacts

During the experiment, 136 state changes for the fridge (9 per participant), 143 for the cupboard (9.5 per participant),

Category	Sound Classe	Sound Nb.	Mean SNR (dB)	Mean length (ms)	Total length (s)
Human sounds:		**36**	**12.05**	**100.8**	**3.35**
	Cough	8	14.6	79	0.6
	Fart	1	13	74	0.01
	Gargling	1	18	304	0.3
	Hand Snapping	1	9	68	0.01
	Mouth	2	10	41	0.01
	Sigh	12	11	69	0.8
	Song	1	5	692	0.7
	Throat Roughing	1	6	16	0.02
	Whistle	5	7.2	126	0.6
	Wiping	4	19.5	76	0.3
Outdoor sounds:		**45**	**9**	**174.4**	**7.85**
	Exterior	24	10	32	0.77
	Helicopter	5	10	807	4.4
	Rain	3	6	114	0.3
	Thunder	13	7.5	208	2.7
Device sounds:		**72**	**8.03**	**208.5**	**15.1**
	Bip	2	8	43	0.08
	Phone ringing	69	8	217	15
	TV	1	10	40	0.04
Water sounds:		**36**	**10.1**	**1756.1**	**63.2**
	Hand Washing	1	5	212	0.2
	Sink Drain	2	14	106	0.2
	Toilet Flushing	20	12	2833	56.6
	Water Flow	13	7	472	6.1
Other sounds:		**395**	**9.5**	**93.9**	**37.1**
	Mixed Sound	164	11	191	31.3
	Unknown	231	8.5	25	5.8
Object handling:		**1302**	**11.9**	**58.6**	**76.3**
	Bag Frisking	2	11.5	86	0.1
	Bed/Sofa	16	10	15	0.2
	Chair Handling	44	10.5	81	3
	Chair	3	9	5	0.01
	Cloth Shaking	5	11	34	0.1
	Creaking	3	8.7	57	0.1
	Dishes Handling	68	8.8	70	4.7
	Door Lock&Shut	278	16.3	93	25
	Drawer Handling	133	12.6	54	7
	Foot Step	76	9	62	4
	Frisking	2	7.5	79	0.1
	Lock/Latch	162	15.6	80	12.9
	Mattress	2	9	6	0.01
	Object Falling	73	11.5	60	4.4
	Objects shocking	420	9	27.6	11.6
	Paper noise	1	8	26	0.03
	Paper/Table	1	5	15	0.01
	Paper	1	5	31	0.03
	Pillow	1	5	2	0
	Rubbing	2	6	10	0.02
	Rumbling	1	10	120	0.1
	Soft Shock	1	7	5	0
	Velcro	7	6.7	38	0.2
Overall sounds except speech		1886	11.2	107.8	203.3
Overall speech		669	11.2	435	291.0
Overall		**2555**	**11.2**	**193.5**	**494.3**

Table 4 Every Day Life Sound Corpus constituted during the experiments of all the subjects and annotated afterwards.

and 91 for the chest of drawers (6 per participant) were recorded. This data is particularly interesting to track furniture usage when preparing a meal or dressing.

6.2.3 Audio

Every audio event was processed on the fly by AUDITHIS and stored on the hard disk. For each one, an XML file was generated, containing the results of the process. These events do not include the ones discarded because of their low SNR (less than 5 dB, threshold chosen empirically). The events were then filtered to remove duplicate instances (same event recorded on different microphones).

During the experiment, 1886 individual sounds and 669 sentences were collected. These periods were manually annotated after the experiment. All the details and description of the corpus are given in Table 4.

The total duration of the audio corpus, including sounds and speech, is 8 min 23 s. This may seem short, but daily living sounds last 0.19s on average. Moreover, the person is alone at home, therefore she rarely speaks (only on the phone). Similarly, few sounds are emitted excepted during particular activities or when the person is moving in the flat.

The mean SNR of each class is between 5 and 15 dB, far less than the in-lab one. This confirms that the health smart home audio data acquired was noisy. Also, the sounds were very diverse, much more than expected in this experimental conditions were participants, though free to perform activities as they wanted, had recommendations to follow.

The speech part of the corpus was recorded in noisy conditions (SNR=11.2dB) with microphones set far from the speaker (between 2 and 4 meters) and was made of phone conversations (no record of the microphone of the telephone) such as *"Allo"*, *"Comment ça va"* or *"A demain"*. No emotional expression was asked from the participants.

According to their origin and nature, sounds were categorised into sounds of daily living classes. A first class is constituted of all the ones generated by the human body. Most of them are of low interest (e.g., clearing throat, gargling, singing). However, whistling and song can be related to the mood while cough and throat roughing may be related to health. The most populated class of sound is the one related to the object and furniture handling (e.g., door shutting, drawer handling, rummaging through a bag, etc.). The distribution is highly unbalanced and it is unclear how these

sounds can be related to health status or distress situation. However, they contribute to the recognition of activities of daily living which are essential to monitor the person's activity. Related to this class, though different, were sounds provoked by devices, such as the phone.

The most surprising class was the sounds coming from the exterior of the flat but within the building (elevator, noise in the corridor, etc.) and exterior to the building (helicopter, rain, etc. outside). This flat has poor soundproofing (as it can be the case for many homes) and we did not prevent participants any action. Thus, some of them opened the window, which was particularly annoying (the helicopter spot of the hospital is at short distance). Furthermore, one of the recordings was realized during rain and thunder which dramatically increased the number of sounds.

It is common, in daily living, for a person, to generate more than one sound at a time. Consequently, a large number of mixed sounds were recorded (e.g. mixing of foot step, door closing and locker). Unclassifiable sounds were also numerous and mainly due to situations in which video was not enough to mark up, without doubts, the noise occurring on the channel. Even for a human, the context in which a sound occurs is often essential to its classification [31].

Despite the length of the experience, the number of sounds that were recorded is low and highly unbalanced for most classes. Thus, the record of a sufficient number of sounds needed for statistical analysis method will be a hard task. This corpus represents a very rare resource of original in home sounds.

6.2.4 Video

75 video recordings were collected (5 per participants, one per room). One video is devoted to the recording of the door of the bathroom (in order to know with kind of activity is happening). This information can be retrieve elsewhere. The four other videos are brought together in order to make one only video for the whole session. These videos were used to mark up the data.

Some other work use video recording as an extra sensor. For this dataset, it seems very difficult to do that. In the condition of this smart home, the installation that we made has to be minimal and efficient (for a possible deployment after). As we chose, for matters of privacy, to remove video from possible sensors, the images were used only to annotate the corpus. Considering this use at that time, we used webcameras (Creative Live Cam Voice) with large angle of reception (89^o) for video recording. These webcameras were installed on USB 2 ports of a PC, with three camera for one computer, so the limitation of bandwidth rapidly appeared and forced us to use a resolution of 320x256 with a low frame rate of 15fps. One camera on each computer (bedroom and corridor) also records the sounds that are emitted in the room. For use of annotation it is largely sufficient. Moreover, we decide also to compress on the fly the video, using an MPEG4-based algorithm (XviD) for storage on server purposes (with sound encoded in MP3 at 16kHz mono). This compression adds noise in the image (lossy compression) that make also harder to use image processing algorithms. VirtualDub is used to record these videos and a filter of this software is responsible for the time-stamping of all the frames.

6.2.5 Wearable Sensor

ACTIM6D has created one file for each subject. Unfortunately, for two of the participants, data were corrupted and then unusable. For the others, it has been synchronized with the video data using a predefined movement and then processed by the algorithm (detection of postural transitions and walking periods [12]).

7 Annotation Scheme

In this study, to make the analysis of the user's activities and distress situations possible, three kinds of information were annotated: user's interactions with object, user's utterances and user's states. User's interactions with objects concern the usage of household appliances such as the fridge or the cupboard as well as the noise made while manipulating an object (e.g., dishes). User's utterances concern the speech that the participant articulated. User's state corresponds to the activity the person is being performing or the place in which s/he is located. These kinds of information are quite different in nature but we chose to represent them using only two temporal types : State and Instant. State is represented as a temporal interval and captures processes (e.g., activities) or devices states (e.g., fridge open). We chose to represent user's interactions with devices as states to embed both contextual information (e.g., in kitchen while the fridge is open) and events (e.g. the fridge changed state at 15:32 so there was an interaction). Instant is represented by a single date and concerns only audio information which is seen as transient with respect to the other kinds of information.

A unique schema was designed because of the homogeneity among the annotation types. The following annotation types were defined which are summed up in Table 5: location of the person, activity, chest of drawers door, cupboard door, fridge door, person's posture, sounds of daily living and speech. For the doors state, the content of the annotations is a simple text flag, '1', that indicates an interval of time where the door is open. For the other annotations a pair 'attribute=value' was used as it provides more flexibility to define other attributes for the same annotation type. For instance, when annotating an activity, besides the attribute-value pair, we can include a short comment about the performed activity which could be useful to assert that there is

Annotation Type	Temporal	Description Type	Domain Value	Content Type
Localization	State	The room in which the person is	Bedroom, living-room, hall, kitchen,bathroom	Text
Activity	State	ADL currently performed	Sleeping, resting, dressing, feeding, Eliminating, hygiene, communicating	Text
Drawer door	State	If door is open or not	1, 0	Text
Cupboard door	State	If door is open or not	1, 0	Text
Fridge door	State	If door is open or not	1, 0	Text
Posture	State	The position of the person's body	Standing, lying, sitting	Text
Sound	Instant	Non-speech sounds emitted by a user or an object	See Table 4 for an exhaustive list	Text
Speech	Instant	Speech utterance of the user	The transcript of the speech utterance	Text

Table 5 Annotation Types

ambiguity or doubt in the information. Metadata containing the creation date and author for each annotation is saved automatically. If necessary, doors state can be easily translated into the 'attribute=value' pair, since their content is simple text. It is worth noting that each annotation type can take only one value in the domain (e.g., no parallel activity, no fuzzy location).

To proceed with the annotations, several tools have been tested, among them we can mention Observer XT, ANVIL, ELAN, and Advene[4]. All these software offer many features to mark-up audio visual data, such as ways of navigating through the recordings, support for different formats, and organisation of the annotations in layers. Advene was chosen because it can deal with large videos without any lost of performance (which was not the case with the other tools). This is possible because Advene does not embed a specific video player but makes use of a wide range of media players that can be called from within the software. Advene organises annotations elements — such as type of annotations, annotations, and relationships — under schemes defined by the user. Some other features of Advene that are worth highlighting are the possibility of specifying relationships among the annotations (e.g., marking up causal links or associations) which provides a semantically rich media description, and the abilities to execute queries on the annotations through the TALES component. A snapshot of Advene with our video data is presented on Figure 4. In the top of the figure, the menu is presented, below it is the video being annotated. At the bottom of the figure, the annotation panel shows the different annotation variables (one per line) and their instances along the temporal dimension. For binary variables (e.g., fridge state), an interval represent the period during which it is open. For non binary state variables (e.g., posture), the value of the instance is displayed within the bar. When the player plays, a dynamic vertical red line in the annotation panel is synchronised with the player.

For the annotation of the non audio data, the procedure was the following. Two annotators marked up the data: the first one marked-up entirely the video and the second one used the input of the first to correct them. The video employed for the annotation, as shown in the figure 4, resulted from a conversion into a mosaic-form of the videos recorded from cameras in the most descriptive four locations of the flat. The top left is the kitchen, the bottom left is the kitchen and the corridor, the top right is the living-room on the left and the bedroom on the right and finally the bottom right represents another angle of the living room. These views were enough to analyse what was going on in every point of the flat.

To mark-up the numerous sounds collected in the smart home, none of the current annotation software showed any advantages. Indeed, the duration of each sounds is too small and the number of audio channels too high (seven) to be properly managed by a resource consuming annotator. Even the Praat software did not revealed to be adequate. We thus used the AUDITHIS software [38] to detect automatically sound event occurrences and to perform a preliminary classification. Then, we developed our own annotator software in the Python language that played each sound while displaying the video in the context of this sound, and that proposed to either keep the AUDITHIS annotation or select another one in a list. All sounds were revisited by an annotator using this software and about 2500 sounds and speech utterances were annotated in this way. While this procedure might led to missing events (i.e., not detected by AUDITHIS) this made it possible to collect automatically numerous sound information such as estimated Signal-to-Noise Ratio, duration, location of the microphone, etc. Moreover, only sounds below 5dB (i.e., signal energy less than about 3 times the one of the background noise) were discarded keeping only the sounds the most reasonably analysable by machine or human.

8 Experiments undertaken with the corpora

The aim of this part is to sum up briefly some experimentations that we made using these datasets. Other researchers (or even us in future publications) could use this dataset for different purposes (especially related to home automation

[4] http://liris.cnrs.fr/advene/

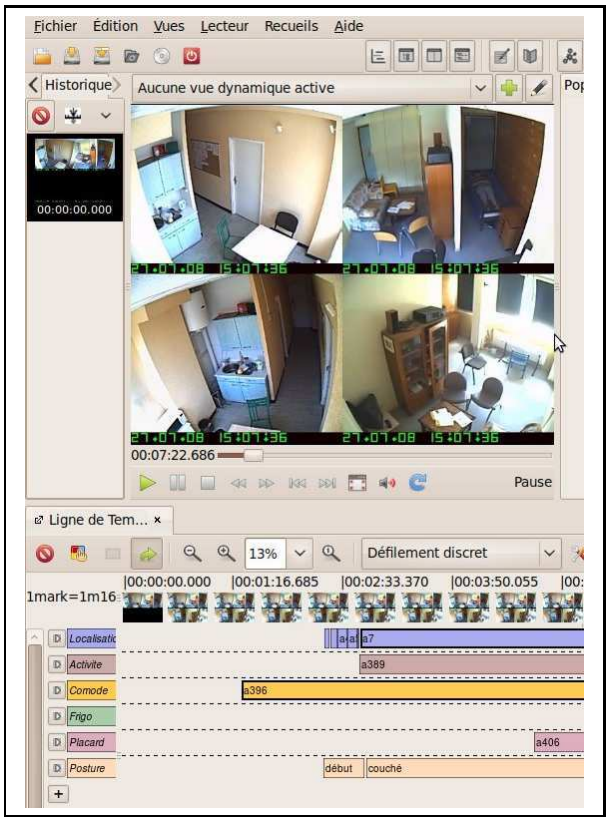

Fig. 4 Snapshot of Advene

system adaptation), so the use of the dataset is not limited to it. The only aim of this section is to present some works done with it (previously published). These works, in case of uses of the dataset, could be some comparison elements for the performances of other developed algorithms.

8.1 Distress call recognition

In this experience, described in detail in [39], the distress call corpus was used to test the recognition of distress keywords in speech occurrences. From the different occurrences, the test was to identify at least one keyword that would permit to identify a distress call. As described in Section 6.1, each audio event could be captured by several microphones in the flat. If the SNR of each record was sufficiently high (at least 5 dB) this record was classified either as a speech utterance or a none-speech sound by a GMM (Gaussian Mixture Model) classifier trained on a separate corpus. If the sound event was classified as speech, it was analysed by an ASR system which output the five highest transcription hypotheses. As many simultaneous detections occurred, each of which generating several transcription hypotheses, a decision stage was introduced. This stage selected the most probable transcription hypothesis. This selection was implemented using four different voting methods:

1. the best hypothesis of the ASR system fed by the speech event with the highest SNR ;
2. a weighting of the three best hypotheses of the ASR system fed by the speech event with the highest SNR ;
3. the same weighting strategy as method 2 but including the three best hypotheses of all the ASR systems fed by the simultaneous speech events with SNR above 80% of the highest SNR ;
4. and the same method as method 1 but using a small specialised language model.

The recognition results of these four classifications are presented in table 6. It can be noticed that distress call recognition is a very difficult task (from about 30 to 60 % error rate). The specialisation of the language model improves the results but they stay unsatisfactory. The most prevalent problem was at the ASR stage. The system was very sensitive to pronunciation defaults and also to artefacts. This shows that speech recognition alone might be insufficient to detect distress calls and that other information such as prosody or context should be integrated in the decision. In a context of smart home, such system could also be specialized with a specific training for each individual (to tune the generic models). When setting up the smart home, it could be done quickly and easily with a little session of recording. It was not our first choice (that was to investigate the performances of a generic model that do not require any further configuration when deployed), but it would be a choice to reduce and optimize the performances of the distress detection using sound.

8.2 Multimodal Recognition of Activity of Daily Living

The first use of the multimodal daily living corpus (cf. Section 6.2) was for activity recognition from non visual sensors [14]. In this work, we investigated the Support Vector Machine classification method to recognize the current activity of the person, from the different sensor data. For this work, a sliding windows approach was used to analyse each temporal chunk of data (here 3-minute chunks). In each chunk was described by a vector of attributes which were chosen to be as representative as possible of the dynamic of the data. A study of the impact of the modality considered for this activity recognition task was also done in [34,38]. The result

	Method 1	Method 2	Method 3	Method 4
Distress	60.1	63.1	54.8	29.5
Normal	10.4	10.0	9.6	4.0
Global	33.4	34.5	30.5	15.6

Table 6 Classification error rate (in percent) for the distress call recognition.

of the classification of seven different activities using this method is given in the first column of Table 7.

This method gave interesting results but it is well known that activity recognition should consider context to reach higher performances [8]. Thus, we integrated prior knowledge in the process. The context considered was the location of the person (i.e., the spatial context) and the time of the day (i.e., the temporal context). Indeed, most of the activities that were performed are often attached to a room (e.g., eating in the kitchen) or to a moment of the day (e.g., eating during lunchtime). From these assumptions, we performed different classifications that took into account these two parameters of the activity and a last one that aggregated the results of the two classifications with prior with the first one (without any). Results of these experiments are presented in [13] and summed up in Table 7.

The poor results of the spatial context was surprising and we investigated the content of the dataset. We discovered that the PIR sensors had low sensitivity and missed some detections. This had a strong impact in the recognition of some activity as one missed detection may lead to believe that the person is in the kitchen while she is asleep for a while.

To solve this problem, we developed a localisation method using several of the sensors of the corpus (from infrared sensors to microphones) to fuse information coming from different sources taking ambiguity into account. The results are presented in [4]. It shows that 88.9% of the location detections are correct considering the infrared sensors and 26% considering the doors contacts. Using these two sensors and also the microphones the ration of correct classification is improved by 2%. This small improvement can be explained because a long event that was missed did not include the sound instances (sleeping). In another smart home environment where infrared sensors were even less performing the interest of fusing the multiple sources including the audio one was much more evident than in this case.

Activity	Generic	Spatial	Temporal	Hybrid
Sleeping	98.0	69.4	100.0	98.0
Resting	78.1	78.1	83.6	78.1
Dress/undress	80.0	73.3	86.7	80.0
Eating	97.8	97.8	97.8	97.8
Toilet use	81.3	87.5	87.5	87.5
Hygiene	71.4	71.4	78.6	71.4
Communication	80.0	65.0	85.0	80.0
Global	86.2	78.9	90.1	86.8

Table 7 Correct classification rate (in percent) of 7 activities with the daily living corpus. All classification are made using a SVM method with a Gaussian kernel. Generic column give the results without any prior consideration, Spatial and Temporal consider knowledge on the room in which an activity can be done or a temporal frame, and Hybrid consider both.

The few experiments presented in this section show that the datasets constituted represent a very good material to develop and evaluate sensing technology in a very realistic and challenging setting.

9 Discussion

The datasets introduced in this paper are valuable for the community interested in smart home and multimodal interfaces. Their usefulness for validation and model acquisition was tested in numerous studies related to context recognition and domotic interfaces [13,4,34,38,14,39]. Moreover, these datasets are amongst the few released that consider seriously multichannel audio as primary source of information. They can thus serve as a stepping stone for the development of audio based interfaces.

Regarding the naturalness of the data, this was ensured by the few recommendations that were given to the participant. Indeed, in the Multimodal Daily Living corpus participants were only told to perform some activities but not how and when these activities should be performed. For instance, they had to prepare a breakfast and eat it, but it was never told how to prepare it and what to eat. By this way many possibilities were offered to the participant to reflect, as much as possible, the reality that the dwellers could encounter at home. It could be argue that this naturalness comes with reduced control, but as shown in Section 6.2, such a rich corpus could not have been acquired with a too constrained protocol. Many challenges in this domain had been drawn from the data analysis [40]. Regarding, the Distress Calls corpus, the participant had to read predefined sentences and were not at any stage in real distress, they could act it. The naturalness of this dataset can thus be discussed. However, the recording conditions were not fully controlled as any participant could choose to turn her back to the microphones (actually not shown to the participants). Moreover, to the best of our knowledge, we are not aware of any other data set of distress calls (acted or not) recorded in a domestic environment.

Regarding the completeness of the data, it must be emphasised that a strong orientation was to use audio sensors reasonably (in term of cost) distributed over the flat (about 2 per rooms). These are the only sensors recorded in the Distress Call corpus and it could be argue that video information could have been useful for post analysis. However, the Multimodal Daily Living contains a much larger set of sensors but chosen to be as simple as possible ones (PID, door contacts, temperature), and having no intersection in information (e.g., audio and video). Moreover, a wearable sensors were used to give even more information about the participant. All these data were annotated at a macro level. Posture, activity, speech, object usages etc. are timely marked up, but this may not be enough for deeper analysis. This is why all the raw data are delivered. For instance, the actimetry is not

released only with the result of the processing that could contains the position of the person, but also the raw data of accelerometers and magnetometers. Any other researches could test it with their own algorithms and assess the results using the annotations or video recordings. Similarly, the audio data can be processed again to compare the results of automatic multisource speech recognition systems.

These datasets can be particularly useful for the study of human behaviour and the development of numerous kinds of interface. For the datasets related to speech, the main use is to interact with the inhabitant to detect the need of help and vocal commands. For instance, after a first recognition of a distress-oriented sentence, it is then possible to launch a process that will try to get more precision from the inhabitant or from the situation to get (through the communication modalities set-up in the home), and then if no correct answer is received or if a real distress situation is detected, to call for emergency. In the non-distress situation, speech processing and audio processing can be used directly for voice command or context assessment (e.g., watching TV, manipulating dishes...). Regarding the multimodal dataset, although it was acquired for activity recognition from non visual sensors, it can be used for many purposes such as scene recognition, inhabitant tracking, emotion recognition, etc. In the project context the dataset was acquired as a first step towards activity evaluation in order to recognize when a person (often an elderly person) is loosing autonomy by failing to realise some daily task. This could then be used either for coaching feedbacks or for support to physicians to assess the state of the person.

Perhaps one of the most interesting application of these datasets is the development of intelligent home automation interfaces that can understand the context in which the person is interacting with the home. For instance, in the SWEET-HOME project [35], voice command is designed to be processed according to the context in which the inhabitant is. As an example, if a user asked for the light to be turned on, the current activity of the user is useful to determine which lamp must be used and at which intensity. If the person is cleaning the house, full intensity of the ceiling lamp will be used whereas if the person is reading on her bed a gentle illumination from the bedside lamp is required. This will make the home automation interfaces more reliable and more friendly to use, given that the decision is made knowing what the user is currently doing in the home bringing mode information about her intention.

10 Conclusion

Despite a growing research interest in smart homes and ambient intelligence only a few number of datasets have been made available to the community. These datasets are crucial to analyse human behaviour, acquire statistical models and validate new ambient technology for activity recognition, distress recognition or speech interface. To be of interest to the largest community, datasets should contain at least some classical home automation sensors, a good variety of users and be as close as possible to realistic situations. In this paper, two corpora are presented both of them having the audio modality as primary channel of sensing, a modality which was generally neglected in other corpora. The first one is a speech only dataset made of 450 sentences uttered by 10 speakers (distress calls, home automation orders and usual conversation). The second one is a multimodal dataset, acquired on 15 subjects in daily living conditions in a fully equipped health smart home. It has been annotated using video cameras and audio records to mark-up activities of daily living periods, interactions with object, speech interactions, sounds of daily living, user's position and others information. This dataset was used a number of time for pattern recognition and classification purposes. Theses datasets are now available on-line [17] to allow the research community with a wide range of interests to use it for model acquisition, analysis or algorithm validation from pattern recognition purpose to the design of intelligent vocal home automation interfaces.

References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. **43**, 1–43 (2011)
2. Badii, A., Boudy, J.: CompanionAble - integrated cognitive assistive & domotic companion robotic systems for ability & security. In: Proceedings of the First Congress of the Société Française des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG'09), pp. 18–20. Troyes (2009)
3. Callejas, Z., López-Cózar, R.: Designing smart home interfaces for the elderly. SIGACCESS Newsletter **95** (2009)
4. Chahuara, P., Portet, F., Vacher, M.: Fusion of audio and temporal multimodal data by spreading activation for dweller localisation in a smart home. In: STAMI, Space, Time and Ambient Intelligence, pp. 17–22. Barcelona, Spain (2011)
5. Chan, M., Estève, D., Escriba, C., Campo, E.: A review of smart homes- present state and future challenges. Computer Methods and Programs in Biomedicine **91**(1), 55–81 (2008)
6. Cook, D.J., Schmitter-Edgecombe, M.: Assessing the quality of activities in a smart environment. Methods of Information in Medicine **48**(5), 480–485 (2009). URL `http://www.ncbi.nlm.nih.gov/pubmed/19448886`
7. Cornet, G., Carré, M.: Technologies pour le soin, l'autonomie et le lien social des personnes âgées : quoi de neuf ? Gérontologie et société **126**, 113–128 (2008)
8. Coutaz, J., Crowley, J.L., Dobson, S., Garlan, D.: Context is key. Communications of the ACM **48**(3), 49–53 (2005)
9. Demiris, G., Rantz, M., Aud, M., Marek, K., Tyrer, H., Skubic, M., Hussam, A.: Older adults' attitudes towards and perceptions of "smart home" technologies: a pilot study. Medical Informatics and the Internet in Medicine **29**(2), 87–94 (2004)
10. Dey, A.K.: Understanding and using context. Personal and Ubiquitous Computing **5**(1), 4–7 (2001)
11. Filho, G., Moir, T.J.: 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 (2010)
12. Fleury, A., Noury, N., Vacher, M.: A wavelet-based pattern recognition algorithm to classify postural transition in humans. In: 17th European Signal Processing Conference (EUSIPCO 2009), pp. 2047 – 2051. Glasgow, Scotland (2009)
13. Fleury, A., Noury, N., Vacher, M.: Improving supervised classification of activities of daily living using prior knowledge. International Journal of E-Health and Medical Communications **2**(1), 17 – 34 (2011)
14. Fleury, A., Vacher, M., Noury, N.: SVM-based multi-modal classification of activities of daily living in health smart homes: Sensors, algorithms and first experimental results. IEEE Transactions on Information Technologies in Biomedicine **14**(2), 274–283 (2010)
15. Gödde, F., Möller, S., Engelbrecht, K.P., Kühnel, C., Schleicher, R., Naumann, A., Wolters, M.: Study of a speech-based smart home system with older users. In: International Workshop on Intelligent User Interfaces for Ambient Assisted Living, pp. 17–22 (2008)
16. Hamill, M., Young, V., Boger, J., Mihailidis, A.: Development of an automated speech recognition interface for personal emergency response system. Journal of NeuroEngineering and Rehabilitation **26**(6) (2009)
17. Health Smart Home Datasets: URL `http://getalp.imag.fr/HISData`
18. Hornbrook, M., Stevens, V., Wingfield, D., Hollis, J., Greenlick, M., Ory, M.: Preventing falls among community-dwelling older persons: results from a randomized trial. Gerontologist **34**(1), 16–23 (1994)
19. Intille, S., Larson, K., Tapia, E., Beaudin, J., Kaushik, P., Nawyn, J., Rockinson, R.: Using a Live-In Laboratory for Ubiquitous Computing Research. In: K. Fishkin, B. Schiele, P. Nixon, A. Quigley (eds.) Pervasive Computing, *Lecture Notes in Computer Science*, vol. 3968, chap. 22, pp. 349–365. Springer Berlin / Heidelberg, Berlin, Heidelberg (2006). DOI 10.1007/11748625_22. URL `http://dx.doi.org/10.1007/11748625_22`
20. Kang, M.S., Kim, K.M., Kim, H.C.: A questionnaire study for the design of smart homes for the elderly. In: Proceedings of Healthcom 2006, pp. 265–268 (2006)
21. van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: UbiComp '08: Proceedings of the 10th international conference on Ubiquitous computing, pp. 1–9. ACM, New York, NY, USA (2008). DOI http://doi.acm.org/10.1145/1409635.1409637
22. Katz, S., Akpom, C.: A measure of primary sociobiological functions. International Journal of Health Services **6**(3), 493–508 (1976)
23. Koskela, T., Väänänen-Vainio-Mattila, K.: Evolution towards smart home environments: empirical evaluation of three user interfaces. Personal and Ubiquitous Computing **8**, 234–240 (2004)
24. Lawton, M., Brody, E.: Assesment of older people: self-maintaining and instrumental activities of daily living. Gerontologist **9**, 179–186 (1969)
25. Le Bellego, G., Noury, N., Virone, G., Mousseau, M., Demongeot, J.: A model for the measurement of patient activity in a hospital suite. IEEE Transactions on Information Technology in Biomedicine **10**(1), 92 – 99 (2006)
26. Lecouteux, B., Vacher, M., Portet, F.: Distant speech recognition in a smart home: Comparison of several multisource asrs in realistic conditions. In: Interspeech 2011, pp. 2273–2276. Florence, Italy (2011)
27. López-Cózar, R., Callejas, Z.: Multimodal dialogue for ambient intelligence and smart environments. In: H. Nakashima, H. Aghajan, J.C. Augusto (eds.) Handbook of Ambient Intelligence and Smart Environments, pp. 559–579. Springer US (2010)
28. Mäyrä, F., Soronen, A., Vanhala, J., Mikkonen, J., Zakrzewski, M., Koskinen, I., Kuusela, K.: Probing a proactive home: Challenges in researching and designing everyday smart environments. Human Technology **2**, 158–186 (2006)
29. Meyer, S., Rakotonirainy, A.: A survey of research on context-aware homes. In: ACSW Frontiers, pp. 159–168 (2003)
30. Moncrieff, S., Venkatesh, S., West, G.A.W.: Dynamic privacy in a smart house environment. In: IEEE Multimedia and Expo, pp. 2034–2037 (2007)
31. Niessen, M.E., van Maanen, L., Andringa, T.C.: Disambiguating sounds through context. In: Proceedings of the second IEEE International Conference on Semantic Computing, ICSC2008, pp. 88–95. IEEE Computer Society (2008)
32. Noury, N., Hervé, T., Rialle, V., Virone, G., Mercier, E.: Monitoring behavior in home using a smart fall sensor and position sensors. In: IEEE-EMBS Microtechnologies in Medicine & Biology, pp. 607–610 (2000)
33. Noury, N., Poujaud, J., Fleury, A., Nocua, R., Haddidi, T., Rumeau, P.: Activity Recognition in Pervasive Intelligent Environments, *Atlantis Ambient and Pervasive Intelligence*, vol. 4, chap. Smart Sweet Home: a pervasive environment for sensing our daily activity, p. 328. Atlantis Press (2011)
34. Portet, F., Fleury, A., Vacher, M., Noury, N.: 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), pp. 63–64. Verona, Italy (2009)
35. Portet, F., Vacher, M., Golanski, C., Roux, C., Meillon, B.: Design and evaluation of a smart home voice interface for the elderly - acceptability and objection aspects. Personal and Ubiquitous Computing pp. 1–30 (2011). DOI http://dx.doi.org/10.1007/s00779-011-0470-5. (in press)
36. Rialle, V., Rumeau, P., Cornet, G., Franco, A.: Les gérontechnologies : au cœur de l'innovation hospitalière et médico-sociale. Techniques hospitalières **703**, 53–58 (2007)
37. Sharkey, A., Sharkey, N.: Granny and the robots: ethical issues in robot care for the elderly. Ethics and Information Technology pp. 1–14 (in press)
38. Vacher, M., Fleury, A., Portet, F., Serignat, J.F., Noury, N.: New Developments in Biomedical Engineering, chap. Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living, pp. 645 – 673. Intech Book (2010)
39. Vacher, M., Fleury, A., Serignat, J.F., Noury, N., Glasson, H.: Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment. In: Proceedings of Interspeech 2008, pp. 496–499. Brisbane, Australia (2008)
40. Vacher, M., Portet, F., Fleury, A., Noury, N.: Development of audio sensing technology for ambient assisted living: Applications and challenges. International Journal of E-Health and Medical Communications **2**(1), 35–54 (2011)
41. Van-Thinh, V., Bremond, F., Thonnat, M.: Automatic video interpretation: a novel algorithm for temporal scenario recognition. In: Proceedings of the 18th international joint conference on Artificial intelligence, pp. 1295–1300. Morgan Kaufmann Publishers Inc., Acapulco, Mexico (2003)
42. Weiser, M.: The computer for the 21st century. Scientific American **265**(3), 66–75 (1991)
43. Youngblood, G.M., Cook, D.J.: Data mining for hierarchical model creation. IEEE Transactions on Systems, Man, and Cybernetics, Part C **37**(4), 561–572 (2007)
44. Ziefle, M., Wilkowska, W.: Technology acceptability for medical assistance. In: 4th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth). Munich, Germany (2010)
45. Zouba, N., Bremond, F., Thonnat, M., Anfosso, A., Pascual, E., Mallea, P., Mailland, V., Guerin, O.: A computer system to monitor older adults at home: preliminary results. Gerontechnology Journal **8**(3), 129–139 (2009)