HAL
open science

# The Sweet-Home speech and multimodal corpus for home automation interaction

Michel Vacher, Benjamin Lecouteux, Pedro Chahuara, François Portet, Brigitte Meillon, Nicolas Bonnefond

## To cite this version:

Michel Vacher, Benjamin Lecouteux, Pedro Chahuara, François Portet, Brigitte Meillon, et al.. The Sweet-Home speech and multimodal corpus for home automation interaction. The 9th edition of the Language Resources and Evaluation Conference (LREC), May 2014, Reykjavik, Iceland. pp.4499-4506. hal-00953006

**HAL Id: hal-00953006**

**https://hal.science/hal-00953006**

Submitted on 4 Jun 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# The Sweet-Home speech and multimodal corpus for home automation interaction

**M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, N. Bonnefond**

Laboratoire d'Informatique de Grenoble, CNRS UMR 5217, Université de Grenoble
41 rue Mathématiques, BP 53, 38041 Grenoble cedex9, France
Michel.Vacher@imag.fr, Benjamin.Lecouteux@imag.fr, Pedro.Chahuara@imag.fr
Francois.Portet@imag.fr, Brigitte.Meillon@imag.fr, Nicolas.Bonnefond@imag.fr

**Abstract**

Ambient Assisted Living aims at enhancing the quality of life of older and disabled people at home thanks to Smart Homes and Home Automation. However, many studies do not include tests in real settings, because data collection in this domain is very expensive and challenging and because of the few available data sets. The SWEET-HOME multimodal corpus is a dataset recorded in realistic conditions in DOMUS, a fully equipped Smart Home with microphones and home automation sensors, in which participants performed Activities of Daily living (ADL). This corpus is made of a multimodal subset, a French home automation speech subset recorded in Distant Speech conditions, and two interaction subsets, the first one being recorded by 16 persons without disabilities and the second one by 6 seniors and 5 visually impaired people. This corpus was used in studies related to ADL recognition, context aware interaction and distant speech recognition applied to home automation controled through voice.

**Keywords:** multimodal data acquisition, natural language and multimodal interactions, voice-activated home automation

## 1. Introduction

The goal of the Ambient Assisted Living (AAL) domain is to foster the emergence of ICT-based solutions to enhance the quality of life of older and disabled people at home, at work and in the community, thus increasing the quality of life, autonomy, participation in social life of elderly people, and reducing the costs of health and social care. Smart spaces, such as smart homes, are emerging as a way to fulfil this goal (Chan et al., 2008). Smart homes can increase autonomy by enabling natural control of the people's environment through home automation, can support health maintenance by monitoring the person's health and behaviour through sensors, can enhance security by detecting distress situations such as fall.

Typical processing undertaken as part of smart home systems includes, automatic location of the dwellers, activity recognition, Automatic Speech Recognition (ASR), dialogue, context-aware decision, sound analysis, behaviour analysis, etc. These tasks ask for a large amount of resources in order to acquire the most accurate models. However, one of the main problems that impede such research is lack of big amount of annotated data. The acquisition of datasets in smart homes is highly expensive both in terms of material and of human resources. Nowadays most of the available corpora recorded in smart spaces are outcomes of research projects (see for instance BoxLab[1] or the MavHome website[2]). For instance, in the context of meeting room analysis (Mccowan, 2003) or of home automation order recognition from distant speech (Cooke et al., 2006; Christensen et al., 2010). One drawback of these datasets is that they often do not include continuous recording of the audio modality. Either home automation sensor traces are recorded or speech data but rarely both at the same time (with the notable exception of (Fleury et al., 2013)). Hence, it is still difficult to conduct study related to multimodal ASR, distant speech, grammar and emotion using these databases.

This paper introduces the corpus composed of audio and home automation data acquired in a real smart home with French speakers. This campaign was conducted within the SWEET-HOME project aiming at designing a new smart home system based on audio technology (Vacher, 2009). The developed system provides assistance via *natural man-machine interaction* (voice and tactile command) and *security reassurance* by detecting distress situations so that the person can manage, from anywhere in the house, her environment at any time in the most natural way possible.

To assess the acceptance of vocal order in home automation, at the beginning of the project, a qualitative user evaluation was performed (Portet et al., 2013) which has shown that voice based solutions are far better accepted than more intrusive solutions (e.g., video camera). Thus, in accordance with other user studies, audio technology appears to have a great potential to ease daily living for elderly and frail persons. To respect privacy, it must be emphasized that the adopted solution analyses the audio information on the fly and does not store the raw audio signals. Moreover, the speech recognizer is made to recognize only a limited set of predefined sentences which prevents recognition of intimate conversations.

The paper introduces the experimental setting of the experiments conduced to acquired the SWEET-HOME corpus. The content of the corpus is then detailed and the results of some studies are then presented.

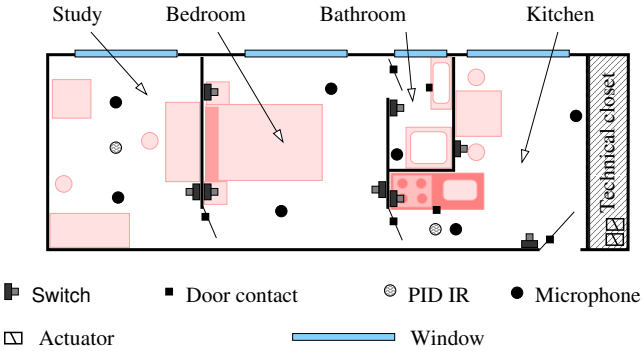
## 2. Smart Home Environment

### 2.1. The DOMUS smart home

To provide data to test and train the different processing stages of the SWEET-HOME system, experiments were con-

[1] http://boxlab.wikispaces.com
[2] http://ailab.wsu.edu/mavhome/research.html

(a) Screenshots of the recording video cameras.




(b) The DOMUS Smart Home with the position of some sensors.

Figure 1: The DOMUS Smart Home.

ducted in the DOMUS smart home of the LIG that was designed to observe users' activities interacting with the ambient intelligence of the environment (see Figure 1). 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.

The data acquisition was performed by recording all possible information sources from the house, including: home automation sensors and microphones. Actuators were driven by an operator (wizard of Oz) or by the intelligent system.

### 2.2. Home Automation Sensors

DOMUS smart apartment is part of the experimentation platform of the LIG laboratory and is dedicated for research projects. DOMUS is fully functional and equipped with sensors, such as energy and water consumption, level of hygrometry, temperature, and actuators able to control lighting, shutters, multimedia diffusion, distributed in the kitchen, the bedroom, the office and the bathroom. A framework has been designed in order to send orders to the different actuators, and to receive the changes of the sensor values (Gallissot et al., 2013).

An observation instrumentation, with cameras, microphones and activity tracking systems, allows to control and supervise experimentations from a control room connected to DOMUS.

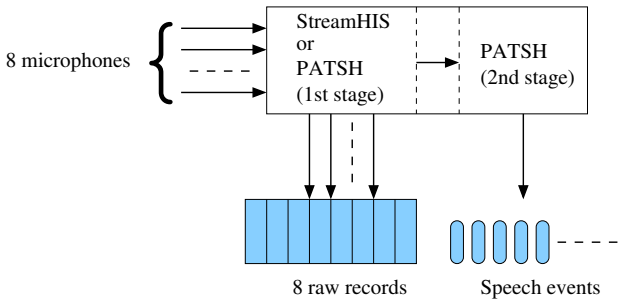



Figure 2: Sound recording.

According to the different research projects, experimentations are conducted with users performing scenarios of daily housework and leisure. Multimodal corpus are produced, synchronized and analyzed in order to evaluate and validate the concerned concept or system.

### 2.3. Audio Recording

The flat also contains 7 radio microphones SENNHEISER ME2 set into the ceiling (2 per room except one for the bathroom) that can be recorded in real-time thanks to a dedicated software able to record simultaneously the audio channels coupled to a National Instrument PCI-6220E multichannel card, the sampling rate was 16 kHz. A 8$^{th}$ channel can be used for a microphone set in front of a noise source (radio or vacuum cleaner).

Two kinds of recording were operated with microphones (see Figure 2). Regarding the first one, a full soundtrack was obtained for each user experiment and may be used for other experiments thanks to StreamHIS (Lecouteux et al., 2011) or PATSH (Vacher et al., 2013c) softwares. These records were then annotated thanks to Transcriber[3], the speech sentences were extracted over the 7 channels and the Signal to Noise Ratio (SNR) was evaluated for each channel.

Regarding the second one, an automatic segmentation was operated by the PATSH software during interaction experiments with the SWEET-HOME system. For each speech event, some information including SNR is available (XML format file). Contrary to the precedent case, the audio data is related to an only channel, the PATSH analyser function has to select the best SNR channel.

## 3. Experiments

Before each experiment, the participant visited the home in order to become familiar with the place and the furniture. The person was asked to read a short text aloud in the smart home in order to adapt the ASR acoustic model to her. To make the acquired data as natural as possible, no instruction was given about how to perform the activity and in which direction to utter sentences. All participants have given an informed and signed consent about the experiment. The acquisition protocol was submitted to the CNIL[4]. CNIL is the French institution that protects personal data and preserves individual liberties.

[3]http://trans.sourceforge.net/
[4]http://www.cnil.fr/english/

### 3.1. Activity of Daily Living

The typical smart homes considered in the study are the one that permit voice based interaction. This kind of smart home can support daily life by taking context sensitive decisions based on the current situation of the user. More specifically, the smart home can be *reactive* to vocal or other commands to make the most adequate action based on context, and can act *pro-actively* by recognising a specific situation in which an action must be made (e.g., for security issue). Therefore, it must be able to recognize the activities of a dweller in order to operate properly in context aware interaction or to provide assistance. Activities can be seen as the atomic elements of a person's day and permit to recognize in which the person is currently, and to detect any deviance from his/her daily routine.

The traditional definition of basic activities, Activities of Daily Living (ADL), was given by (Katz and Akpom, 1976). While ADLs are related to daily human needs, instrumental ADLs (iADLs) are focused on the activities that involves higher level organisation (e.g., handling the phone) or handling of tools (e.g., preparing food). Table 1 presents some examples of these two kinds of activities.

| ADL | iADL |
|---|---|
| Feeding | Using the phone |
| Dressing | Preparing and managing food |
| Hygiene | Doing shopping |
| Locomotion | Doing laundry |

Table 1: ADL and iADL examples

In the context of our application, 7 activities were considered: (1) Sleeping; (2) Resting: managing emails, listening to the radio, reading a magazine; (3) Dressing/undressing; (4) Preparing the meal; (5) Feeding: having a meal; (6) Doing the laundry (dishes and vacuum); (7) Hygiene activity: washing teethes and having a shower, cleaning up the flat using the vacuum; and (7) Communicating: talking with the phone, going out to do the shopping.

Twenty one participants (including 7 women) were asked to perform these activities without any condition on the time spent and or about the way for doing activities. Given the high number of necessary elementary activities and the duration of the experiment (mean: 1h 14mn), a detailed scenario simulating the sequences of a day, as well as a list of sentences to utter during the phone call were given to the participant and explained during the visit of the flat.

Thanks to this experiment, a rich multimodal data set was collected associating speech data recorded in Distant Speech conditions with home automation data. It should be noted that these two types of data are very different from a temporal point of view: the first one is made of analog signals sampled at 16 kHz and the other of network frames transmitted over the home automation network when the person actuated a switch or moved in front of a infra-red detector.

### 3.2. Vocal Order for Home Automation

The project aims at designing a new smart home system based on audio technology to drive home automation thanks to vocal orders. Recent developments have produced significant results and enabled the Automatic Speech Recognizers (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 achieve 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). This is due to different phenomena such as the presence of noise background and reverberation (Vacher et al., 2008) and these issues must be considered in the context of home automation. Moreover, the mouth does not emit sound in the same way for all frequencies, the emitting cone being narrower for high frequencies. These conditions are known as Distant Speech (Wölfel and McDonough, 2009),

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

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

Given that, to the best of our knowledge, no dataset of French utterances of voice commands in a noisy multi-source home exists, we conducted an experiment to acquire a representative speech corpus composed of utterances of not only home automation orders and distress calls, but also colloquial sentences. In order to get more realistic conditions, two types of background noise were considered while the user was speaking: the broadcast news radio and a music (classical) background. These were played in the study through two loud speakers, the 8$^{th}$ channel was used to record the emitted sound thanks to a microphone set in front of one loud speaker. Note that this configuration poses much more challenges to classical blind source separation techniques than when speech and noise sources are artificially linearly mixed.

Voice orders were defined using a very simple grammar as shown on Figure 3. Our previous user study showed that targeted users prefer precise short sentences over more natural long sentences (Portet et al., 2013). Each order belongs to one of three categories: initiate command, stop command and emergency call (see Figure 3). Except for the emergency call, all command starts with a unique keyword that permits to know whether the person is talking to the smart home or not. In the following, we will use *'Nestor'* as key-word:

- set an actuator on: (e.g. `Nestor ferme fenêtre`)
  `key initiateCommand object;`

- stop an actuator: (e.g. `Nestor arrête`)
  `key stopCommand [object];`

- emergency call: (e.g. `au secours`).

The protocol was composed of two steps. During the first step, the participant was asked to go to the study, to close the door and to read a short text of 285 words. This data was used for the adaptation of the acoustic model of the ASR. During the second step, the participant uttered 30 sentences in different rooms in different conditions. The first condition was without additional noise, the second one was with the radio turned on in the study and the third one was with classical music played in the study. Each sequence of 30 sentences was composed by a random selection of 21 home automation orders (9 without initiating keyword), 2 distress calls (e.g., "A l'aide" (help), "Appelez un docteur' (call a doctor)) and 7 casual sentences (e.g., "Bonjour" (Hello), "J'ai bien dormi" (I slept well)). No participant uttered the same sequences. The radio and the music were unique and pre-recorded and were started at a random time for each participant. 23 persons (including 9 women) participated to the experiment. The average age of the participants was 35 years (19-64 min-max). No instruction was given to any participant about how they should speak or in which direction. Consequently, no participant emitted sentences directing their voice to a particular microphone. The distance between the speaker and the closest microphone was about 2 meters. The total duration of the experiment was about 5 hours. During this experiment, only sound data were recorded.

### 3.3. Evaluation of the SWEET-HOME System

The system architecture is presented in Figure 4 and a complete description is available in (Vacher et al., 2013a). The input is composed of the information available from the home automation network and information from 7 microphones. The audio signals are analysed by the PATSH framework in real-time. The signals discriminated as speech are sent to the ASR if they satisfy the following conditions: $SNR \geq 0.1dB$ and $150ms \leq duration \leq 2.8s$. The Speeral tool-kit was used as ASR (Linarès et al., 2007). Speeral relies on an A$\star$ decoder with HMM-based context-dependent acoustic models and trigram language models. At the end, the most probable recognition hypothesis is then sent to the Intelligent Controller.

Thus, information can be provided directly by the user (e.g., voice order) or via environmental sensors (e.g., temperature). The information coming from the home automation system is transmitted on-line to the Intelligent Controller (through several preprocessing steps). This controller captures all streams of data, interprets them and executes the required actions by sending orders to actuators through the home automation network. Moreover, the controller can send alert or information messages to a speech synthesiser in case of emergency or on request of the user ("Beware, the front door is not locked", "The temperature is 21 degrees") (Chahuara et al., 2013).

This experiment aims at evaluating the context awareness performances of the complete SWEET-HOME system in realistic conditions. Two cases were considered: the first one implies users without disabilities and the second one aged or visually impaired persons.

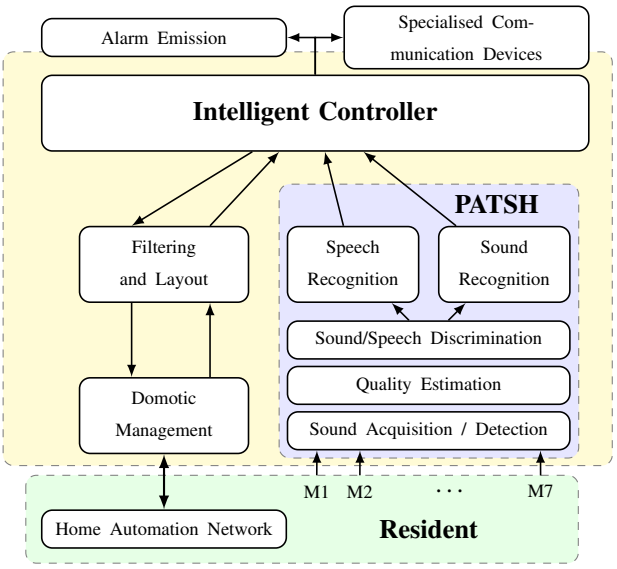



Figure 4: Block diagram of the SWEET-HOME system.

#### 3.3.1. Population without disabilities

To validate the system in realistic conditions, we built scenarios in which every participant was asked to perform the following activities: (1) Sleeping; (2) Resting: listening to the radio; (3) Feeding: preparing and having a meal; and (4) Communicating: having a talk with a remote person thanks to a specialized communication system *e-lio*[5]. Therefore, this experiment allowed us to process realistic and representative audio events in conditions which are directly linked to usual daily living activities. Moreover, to evaluate the decision making, some specific situations were planned in the scenarios.

Each participant had to use vocal orders to make the light on or off, open or close blinds, ask about temperature and ask to call his or her relative. The instruction was given to the participant to repeat the order up to 3 times in case of failure. In that case, a wizard of Oz was used. Sixteen participants (including 7 women) without special needs were asked to perform the scenarios without condition on the duration after a visit of the flat allowing to explain the right way to utter vocal orders and to use the *e-lio* system. Before the experiment, the participant was asked to read a text of 25 short sentences in order to adapt the acoustic models of the ASR for future experiments.

The average age of the participants was 38 years (19-62, min-max) and the experiment lasted between 23min and 48min. The scenario includes at least 15 vocal orders for each participant but more sentences were uttered because of repetitions.

#### 3.3.2. Aged or visually impaired population

The method was also applied in the same context but the protocol was a little simplified and the experiment was cut into 4 independent scenarios:

- The participant is eating his/her breakfast and is preparing to go out. He/she asks the house to turn on

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

| | The Sweet Home Corpus subsets | | | |
|---|---|---|---|---|
| Attributes | Multimodal | Home Automation | Interaction | User Specific |
| # Participants | 21 | 23 | 16 | 11 |
| Age (min-max) | 22-63 | 19-64 | 19-62 | 50-91 |
| Gender | 7F, 14M | 9F, 14M | 7F, 9M | 8F, 3M |
| Ridden text for ASR adaptation | No | Yes (285 words) | Yes (288 words) | Yes (288 words) |
| Duration per channel | 26h | 3h 06mn | 8h 52mn | 4h 39mn |
| Speech (#sentences), in French | 1779 | 5520 (2760 with added noise effect) | 993 | 907 |
| Sounds (#events) | - | - | 3503 | 1866 |
| Home automation traces | yes | no | yes | yes |
| Added noise effect | no | yes (vacuum, TV, radio) | no | no |
| Home Automation orders | no | yes | yes | yes |
| Distress call | yes | yes | yes | yes |
| Colloquial sentence | yes | yes | yes | yes |
| Interaction | no | no | yes | yes |
| Segmented speech | manually | manually | manually and automatically | automatically |
| Transcribed speech | yes | yes | yes | yes |
| Transcribed SNR (at sentence level) | yes | yes | yes | yes |
| Transcribed home automation traces | yes | no | no | no |
| Transcribed sounds | no | no | automatically segmented | automatically segmented |

Table 2: Description of the 4 subsets of the Sweet-Home Corpus

the light, close the blinds or ask for the temperature while doing these activities.

- The participant is coming back from shopping and is going to have a nap. He/she asks the same kind of commands but in this case, a warning situation alerts about the front door not being locked.
- The participant is going to the study to communicate with one relative through the dedicated *e-lio* system. After the communication, the participant simulates a sudden weakness and call for help.
- The participant is waiting in the study for friends going to visit her. She tests various voice orders with the radio, lights and blinds.

At the beginning of the experiment, the participant were asked to read a short text (288 words), this text was used to adapt the acoustical models to the speaker before performing the scenarios. Eleven participants were recruited to perform these scenarios either aged (6 women) or visually impaired (2 women, 3 men). During this experiment, about 4 hours of data were recorded.

# 4. The Acquired Corpus

## 4.1. The ADL annotation schema

The corpus was annotated by means of Advene[6], an annotation tool that allows to organize annotations into different types and schemata. The main goal of this annotation was to set the ground truth for the location of the inhabitant and the activity she performs during the experiments. The state of the devices that gives information about the location and the activity are also part of this annotation. The speech was annotated using Transcriber and the Signal to Noise ratio (SNR) was evaluated for each channel. Table 3 shows the types of annotations that were set for the SWEET-HOME corpus. The apartment is composed of four rooms: a kitchen, a bedroom, a bathroom, and a study, therefore it was possible to consider the following activities for the experiment: eating and preparing a meal, cleaning and tiding, dressing/undressing, talking by the telephone, sleeping, reading and listening the radio, and hygiene actions.
We consider that activities are composed of several actions. For the activities we consider in our study, we have defined a set of actions. Thus, the actions involved in the activity eating and preparing a meal are: eating, drinking, washing dishes. While the actions included in hygiene are: washing hands, taking a shower, and tooth brushing. The others activities are composed of only one action.
In the case of Doors, the annotation do not contain the intervals in time with a certain state, but only the events indicating the door involved and its new state. The same has been applied for the position of the person.

## 4.2. General organisation of the corpus

The SWEET-HOME corpus, detailed in Table 2, is divided in four parts, each of which is presented in its own section.

## 4.3. The *Multimodal* subset: ADL experiment

This subset was recorded to train models for automatic human activity recognition and location. These two types of information are crucial for context aware decision making

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

in smart home. For instance, a vocal command such as "allume la lumière (turn on the light)" cannot be handled properly without the knowledge of the user's location. The characteristics of the corpus are showed in the second column of Table 2. During the experiment, event tracks from the home automation network, audio (7 channels) and video sensors were captured. In total, more than 26 hours of data have been acquired (audio, home automation sensors and videos).

### 4.4. The *French Home Automation Speech* subset

As explained in section 3.2., this subset was recorded to develop robust automatic recognition of voice commands in a smart home in distant conditions (the microphones were not worn but set in the ceiling). Seven audio channels were recorded to acquire a representative multichannel speech corpus composed of utterances of not only home automation orders and distress calls, but also colloquial sentences. The sentences were uttered in different rooms and conditions as described Table 4. The first condition was without noise, the second one was with the radio turned on in the study and the third one was with classical music played in the study. No instruction was given to the participants about how they should speak or in which direction. Each sentence was manually annotated on the best Signal-to-Noise Ratio (SNR) channel using Transcriber.

The third column of Table 2 shows the details of the record. It is composed, for each speaker, of a text of 285 words for acoustic adaptation (36 minutes for 351 sentences in total for the 23 speakers), and of 240 short sentences (2 hours and 30 minutes per channel in total for the 23 speakers) with a total of 5520 sentences overall. In clean condition, 1076 voice commands and 348 distress calls were uttered while they were respectively 489 and 192 in radio background noise and 412 and 205 with music.

### 4.5. The *Interaction subset*

The fourth column of Table 2 summarises the last subset of the SWEET-HOME corpus. In the matter of speech records, a text read aloud is available for each participant. Each of them had to use the grammar (see Figure 3) to utter vocal orders to open or close blinds, ask about temperature, ask to call his or her relative... The scenario included 15 vocal orders for each participant but more sentences were uttered because of repetitions due to errors of the recognition system or to a wrong use of the grammar by the participant.

| Annotation Type | Description | Attribute | Content Type |
|---|---|---|---|
| Localisation | The room in which the person is. | room | text |
| Activity | ADL currently performed. | name | text |
| Action | Part of an activity. | | text |
| Position | If the person is laying, sitting, or standing. | posture | text |
| Material | The kind of material the person manipulates (plastic, glass, paper, etc.). | name | text |
| Doors | The door that is opened/closed. | name | text |
| Electrical device | Use of vacum cleaner or radio. | type | text |

Table 3: Annotation Types

| External noise | Study | Bedroom | Bathroom | Kitchen |
|---|---|---|---|---|
| None | 30 | 30 | 30 | 30 |
| Music (radio) | 30 | 30 | - | - |
| Speech (broadcast news) | 30 | 30 | - | - |

Table 4: Home automation subset: Uttered sentence number as a function of the room

| Speaker ID | All | Sound | Speech | Order | SNR (dB) |
|---|---|---|---|---|---|
| S01 | 213 | 184 | 29 | 20 | 17 |
| S02 | 285 | 212 | 73 | 22 | 19 |
| S03 | 211 | 150 | 61 | 26 | 12 |
| S04 | 302 | 211 | 91 | 19 | 25 |
| S05 | 247 | 100 | 48 | 40 | 20 |
| S06 | 234 | 189 | 45 | 37 | 14 |
| S07 | 289 | 216 | 72 | 21 | 14 |
| S08 | 249 | 190 | 59 | 28 | 14 |
| S09 | 374 | 283 | 91 | 32 | 17 |
| S10 | 216 | 163 | 53 | 26 | 18 |
| S11 | 211 | 155 | 56 | 24 | 15 |
| S12 | 401 | 346 | 55 | 33 | 12 |
| S13 | 225 | 184 | 41 | 40 | 11 |
| S14 | 235 | 173 | 62 | 26 | 17 |
| S15 | 641 | 531 | 111 | 27 | 12 |
| S16 | 262 | 216 | 46 | 22 | 14 |
| All | 4.595 | 3.503 | 993 | 443 | - |

Table 5: Interaction subset: Recorded audio event characteristics

During the scenario, the participant coped with 4 situations: 1) **Feeding:** preparing and having a meal; 2) **Sleeping**; 3) **Communicating:** initiating and having a talk with a relative thanks to the specialised device *e-lio*; 4) **Resting:** listening to the radio.

Globally, 443 vocal orders were recorded, and Table 5 shows for each participant the number of syntactically correct vocal orders or distress sentences and the corresponding SNR. The SNR varies between 11dB and 25dB and is below or equal to 17dB for 12 participants.

### 4.6. The *User specific interaction* subset

To assess the system with targeted users (aged or visually impaired), a user study was set up consisting of semi-directed interviews and sessions in which the user, alone in the flat, had to interact with the SWEET-HOME system following predefined scenarios. The scenario was built from the last one of section 4.5. but simplified and cut into 4 short independent sequences. During the last sequence, the participant was able to use is proper syntax of command, and therefore, not all home automation orders are following the predefined grammar. Recorded data are presented Table 6, SI denotes a visually impaired participant and SA an aged participant. A part of the speech data are related to a talk conversation with the relative thanks to the *e-lio* system.

## 5. Some Studies using this Corpus

### 5.1. ADL Recognition

There is an increased interest in automatic human activity recognition from sensors just as well in Ambient Intelligence domain for context-aware interaction (Coutaz et al., 2005) as in Ambient Assisted Living domain for behaviour monitoring in health applications (Fleury et al., 2010).

| Speaker ID | Age, sex | Scenario duration | Speech | Order | SNR (dB) |
|---|---|---|---|---|---|
| SA01 | 91, F | 24mn | 59 | 33 | 16 |
| SI02 | 66, F | 17mn 49s | 67 | 24 | 14 |
| SI03 | 49, M | 21mn 55s | 53 | 26 | 20 |
| SA04 | 82, F | 29mn 46s | 74 | 26 | 13 |
| SI05 | 66, M | 30mn 37s | 47 | 25 | 19 |
| SA06 | 83, F | 22mn 41s | 65 | 27 | 25 |
| SA07 | 74, F | 35mn 39s | 55 | 26 | 14 |
| SI08 | 64, F | 18mn 20s | 35 | 22 | 21 |
| SA09 | 77, F | 23mn 5s | 46 | 23 | 17 |
| SI10 | 64, M | 24mn 48s | 49 | 20 | 18 |
| SA11 | 80, F | 30mn 19s | 79 | 24 | 23 |
| All | - | 4h 39mn | 629 | 276 | - |

Table 6: User specific interaction subset: Recorded audio event characteristics

The Multimodal subset was used for a study related to ADL recognition from non-visual data (Chahuara et al., 2012). We proposed a procedure that uses raw data from non visual and non wearable sensors in order to create a classification model leveraging logic formal representation and probabilistic inference. The Markov Logic Network method was evaluated and compared with SVM and Naive Bayes, as they have proved to be highly efficient in classification tasks. The overall accuracy achieved with MLN is 85.3%, it is significantly higher than the one obtained with SVM (59.6%) and Naive Bayes (66.1%).

### 5.2. Context Aware Interaction

The SWEET-HOME project addresses the issue of building home automation systems reactive to voice for improved comfort and autonomy at home. Context aware interaction is one of the challenges to take up. Our study focus on the context-aware decision process which uses a dedicated Markov Logic Network approach to benefit from the formal logical representation of domain knowledge as well as the ability to handle uncertain facts inferred from real sensor data (Chahuara et al., 2013). This study used the Interaction corpus. The decision process take into account some inferred variables, the location of the person and his activity. The correct decision rate are give Table 7 for the 4 situations described in section 4.5.

| Situation/Expected utility | Decision rate |
|---|---|
| Situation 1 | 54% |
| Situation 2 | 93% |
| Situation 3 | 73% |
| Situation 4 | 60% |
| All situations | 70% |

Table 7: Correct decision rate

### 5.3. Distant Speech Recognition

The French Home Automation Speech subset was used for studies related to vocal order recognition in Distant Speech conditions (Vacher et al., 2012). A novel version of the Driven Decoding Algorithm (DDA-2) was evaluated. The *a priori* knowledge is obtained after decoding the best SNR channel: for decoding the second best SNR channel, the ASR system is driven by the 3 –*nearest* vocal orders of this hypothesis by using edit distance. This approach presents an important benefice, a strict usage of the grammar may bias toward vocal orders but the projection of vocal orders does not prevent the ASR recognising colloquial sentences. Moreover, an Acoustic Echo Cancellation (AEC) technique was used in the case of localized known sources (radio). The corresponding recall results are given Table 8 for two kinds of noise, classical music(CM) and broadcast news (BN).

Very good results were obtained using DDA in noisy and clean conditions, except for broadcast news. AEC improves greatly the performances in the case of broadcast news but leads to poor performances with music. This might be due to the fact that AEC introduces noise and non-linear distortion. Other methods like Blind Source Separation may be considered for performance improvement.

| Signal | Method | Order recall (%) | Distress recall (%) |
|---|---|---|---|
| Without noise | ASR only | 62.1 | 84.2 |
| Without noise | DDA | 92.7 | 87.2 |
| BN | ASR only | 29.3 | 74.3 |
| BN | DDA | 57.2 | 75.2 |
| BN | AEC+DDA | 83.5 | 81.2 |
| MC | ASR only | 59.0 | 81.6 |
| MC | DDA | 90.6 | 85.2 |
| MC | AEC+DDA | 79.2 | 66.5 |

Table 8: Home automation order and distress detection

### 5.4. Evaluation of a context-aware voice interface for Ambient Assisted Living

The SWEET-HOME system was evaluated in the DOMUS flat with three user groups: seniors, visually impaired and people without special needs as related in section 3. (Vacher et al., 2013c). Regarding the technical aspect, DDA was not implemented in the system and consequently the vocal order error rate was high, 38%, and then the participant had often to repeat the orders. But an other objective was to evaluate the system regarding the user's interest, the accessibility the usefulness and the usability, particularly for the senior and visually impaired people. Hence the evaluation allowed to record their feedback about the system.

None of the people said he had any difficulty in performing the experiment. It is worth emphasizing that aged people preferred a manual interaction because this was quicker, the voice warning was well appreciated. The visually impaired people found that the voice control would be more adequate if it could enable performing more complex tasks than controlling blinds or radio.

## 6. Distribution of the corpus

The multimodal and home automation parts are going to be released for research purpose only on a dedicated website `http://sweet-home-data.imag.fr` (Vacher, 2009). To respect privacy, no videos are released. Indeed, video data were only captured for manual marking up and then not available. The last part of the corpus will be released when the annotation process will be finished.

## 7. Conclusion

Smart Homes, despite to their recent developments, have led to few experimentations in daily living conditions. This paper presents a corpus made of 4 subsets acquired in daily living conditions in a fully equipped and complete smart home. Recorded data consist in home automation traces, sound and speech. The experiments that permitted these records are detailed, including the last one where aged and visually impaired users were implied. Finally, the results of experiments related to human location and activity recognition (Chahuara et al., 2012) and vocal command recognition and distress detection in real-time (Vacher et al., 2013b) are presented. Obtained results are not sufficient for a use in real conditions and future works include new studies related to activity recognition improvement and robust speech recognition in realistic noise conditions.

## 8. Acknowledgements

This work is part of the SWEET-HOME project founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-09-VERS-011). The authors would like to thank the participants who accepted to perform the experiments. Thanks are extended to S. Humblot, S. Meignard, D. Guerin, C. Fontaine, D. Istrate, C. Roux and E. Elias for their support.

## 9. References

P. Chahuara, A. Fleury, F. Portet, and M. Vacher. 2012. Using Markov Logic Network for On-line Activity Recognition from Non-Visual Home Automation Sensors. In *Ambient Intelligence*, volume 7683 of *Lecture Notes in Computer Science*, pages 177–192. Springer (Heidelberg).

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

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

H. Christensen, J. Barker, N. Ma, and P. Green. 2010. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Proceedings of Interspeech 2010*, pages 1918–1921.

M. Cooke, J. Barker, S. Cunningham, and X. Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120:2421–2424.

Joëlle Coutaz, James L. Crowley, Simon Dobson, and David Garlan. 2005. Context is key. *Communications of the ACM*, 48(3):49–53.

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

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

M. Gallissot, J. Caelen, F. Jambon, and B. Meillon. 2013. Une plate-forme usage pour l'intégration de l'informatique ambiante dans l'habitat : Domus. *Technique et Science Informatiques (TSI)*, 32/5:547–574.

S. Katz and C.A. Akpom. 1976. A measure of primary sociobiological functions. *International Journal of Health Services*, 6(3):493–508.

B. Lecouteux, M. Vacher, and F. Portet. 2011. Distant speech recognition for home automation: Preliminary experimental results in a smart home. In *IEEE SPED 2011*, pages 41–50, Brasow, Romania, May 18-21.

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

I. A. Mccowan. 2003. The multichannel overlapping numbers corpus. Idiap resources available online: http://www.cslu.ogi.edu/corpora/monc.pdf.

F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon. 2013. Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17(1):127–144.

M. Vacher, A. Fleury, J.-F. Serignat, N. Noury, and H. Glasson. 2008. Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment. In *Proceedings of Interspeech 2008*, pages 496–499, Brisbane, Australia, Sep. 22-26.

M. Vacher, B. Lecouteux, and F. Portet. 2012. Recognition of Voice Commands by Multisource ASR and Noise Cancellation in a Smart Home Environment. In *EUSIPCO*, pages 1663–1667, aug.

M. Vacher, P. Chahuara, B. Lecouteux, D. Istrate, F. Portet, T. Joubert, M. Sehili, B. Meillon, N. Bonnefond, S. Fabre, C. Roux, and S. Caffiau. 2013a. The SWEET-HOME Project: Audio Technology in Smart Homes to improve Well-being and Reliance. In *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13)*, pages 7298–7301, Osaka, Japan.

M. Vacher, B. Lecouteux, D. Istrate, T. Joubert, F. Portet, M. Sehili, and P. Chahuara. 2013b. Evaluation of a Real-Time Voice Order Recognition System from Multiple Audio Channels in a Home. In *Proceedings of Interspeech 2013*, pages 2062–2064.

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

M. Vacher. 2009. The SWEET-HOME project. `http://sweet-home.imag.fr/`. Accessed: 2014-03-30.

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