Experimental Evaluation of Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home

Michel Vacher, Benjamin Lecouteux, Dan Istrate, Thierry Joubert, François Portet, Mohamed Sehili, Pedro Chahuara

Experimental Evaluation of Speech Recognition Technologies for Voice-based Home Automation Control in a Smart Home

Michel Vacher[1], *Benjamin Lecouteux*[1], *Dan Istrate*[2],
Thierry Joubert[3], *François Portet*[1], *Mohamed Sehili*[2], *Pedro Chahuara*[1]

[1]LIG, UMR5217 UJF/CNRS/Grenoble-INP/UPMF, 38041 Grenoble, France
[2]ESIGETEL, 77210 Avon - France
[3]THEORIS, 75000 Paris - France

{Michel.Vacher,Benjamin.Lecouteux,Pedro.Chahuara, Francois.Portet}@imag.fr
dan.istrate@esigetel.fr, mohamed.sehili@esigetel.fr, thierry.joubert@theoris.fr

Abstract

This paper presents an audio-based interaction technology that lets the user have full control over her home environment and at detecting distress situations for the elderly and frail population. We introduce the PATSH framework which performs real-time recognition of voice commands anywhere in the home and detail its architecture and the state-of-the-art processing technologies it employs. This system was evaluated in a realistic Smart Home with three user groups: seniors, visually impaired people and people with no special needs. Results showed the validity of the PATSH approach and shed light on its usability for people with special needs.

Index Terms: Real-time audio analysis, experimental in-situ evaluation, Smart Home, Ambient Assisted Living

1. Introduction

Due to the demographic change and ageing in developed countries, the number of older persons is steadily increasing. In this situation, the society must find solutions to allow these people to live in their home as comfortably and safely as possible by assisting them in their daily life. This concept, known as Ambient Assisted Living (AAL) aims at anticipating and responding to the special needs of these persons. In this domain, the development of Smart homes and intelligent companions is seen as a promising way of achieving in-home daily assistance [1]. However, given the diverse profiles of the senior population (e.g., low/high technical skill, disabilities, etc.), complex interfaces should be avoided. Nowadays, one of the best interfaces seems to be the speech interface, that makes possible interaction using natural language so that the user does not have to learn complex computing procedures or jargon. Moreover, it is well adapted to people with reduced mobility and to some emergency situations because the user doesn't need to be close to a switch ("hands free" system). Despite all this, very few Smart Home projects have seriously considered speech recognition in their design [2, 3, 4, 5, 6, 7, 8]. Part of this can be attributed to the complexity of setting up this technology in a real environment and to important challenges that still need to be overcome [9].

In order to make in home voice control a success and a benefit for people with special needs, we argue that a complete framework for audio analysis in Smart Home must be designed. This framework should be able to provide real-time response, to analyse concurrently several audio channels, to detect audio events, to filter out noise and to perform robust distant speech recognition. Furthermore, in contrast with current triggered-by-button ASR systems commonly found in smart phone, this voice control should be able to work in an "hand free" manner in case the person is not able to move. Another important aspect is the respect for privacy: the system should not disseminate any raw personal data outside the home without the user's consent. Our approach, called PATSH is a step toward these goals. The originality of the approach is to consider these problems together while they have mostly been studied separately.

To the best of our knowledge, the main trends in audio technology in Smart Homes are related to augmented human machine interaction (e.g., voice command, conversation) and security (mainly fall detection and distress situation recognition). Regarding security, the main application is the fall detection using the signal of a wearable microphone which is often fused with other modalities (e.g., accelerometer) [4, 3]. However, the person is constrained to wear these sensors at all times. To address this constraint, the dialogue system developed by [6] was proposed to replace traditional emergency systems that requires too much change in the lifestyle of the elders. However, the prototype had a limited vocabulary (yes/no dialogue), was not tested with aged users and there is no mention about how the noise was taken into account. Most of the speech related research or industrial projects in AAL are actual highly focused on dialogue to build communicative agent (e.g., see the EU funded Companions or CompanionAble projects or the Semvox system [1]). These systems are often composed of ASR, NLU, Dialogue management and TTS parts supplying the user the ability to communicate with the system in an interactive fashion. However, it is generally the dialogue module (management, modelling, architecture, personalization, etc.) that is the main focus of these projects (e.g., see Companions, OwlSpeak or Jaspis). Moreover, this setting is different from the Smart Home one as the user must be close to the avatar to speak (i.e., not a distant speech setting). In [7], a communicative avatar was designed to interact with a person in a smart office. In this research, enhanced speech recognition is performed using beamforming and a geometric area of recording. But this promising research is still to be tested in a multiroom and multisource realistic home.

Designing and applying speech interfaces in Smart Home to provide *security reassurance* and *natural man-machine interaction* is the aim of the SWEET-HOME[2] project. With respect

[1] http://www.semvox.de
[2] http://sweet-home.imag.fr

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 99–105,
Grenoble, France, 21–22 August, 2013.

to this short state-of-the-art, the project addresses the important issues of distant voice command recognition and sound source identification. The outcomes of this research are of high importance to improve the robustness of the systems mentioned above. In this paper, we introduce the PATSH system which perform the real-time identification of the voice command anywhere in the home. Its architecture and the state-of-the-art processing technologies employed are detailed in Section 2. This system was evaluated in a realistic Smart Home with three user groups: people with no special needs, seniors and, visually impaired people. These experiments are summarised in Section 3. PATSH was used on-line (vs. off-line) during the experiment, these results are analysed in Section 4. The paper finishes with a short outlook of future work.

2. The Audio Analysis System

The SWEET-HOME system is composed of an Intelligent Controller which analyses the streams of data and makes decision based on these. This framework acquires data from sensors and interprets them, by means of IA techniques, to provide contextual information for decision making. The description of this intelligent controller is out of the scope of the paper, the reader is thus referred to [12] for further details. This system uses a two-level ontology to represent the different concepts handled during the processing which also contains SWRL instances to automatise some of the reasoning. An important aspect is the relationship between the knowledge representation and the decision process which uses a dedicated Markov Logic Network approach to benefit from the formal logical definition of decision rules as well as the ability to handle uncertain facts inferred from real data. The location of the inhabitant was determined by the intelligent controller that analysed continuously the data stream of the smart-home (not only audio) and made decisions based on the recognized voice commands and this contextual information.

Therefore, the streams of data are composed of all the usual home automation data sensors (switches, lights, blinds, etc.), multimedia control (uPnP), and the audio events processed in real-time by the multi-channel audio analysis system: PATSH. Indeed, this section describes the overall architecture of PATSH, details the sound/speech discrimination and the ASR part.

2.1. PATSH framework

The global architecture of PATSH is illustrated in Figure 1. The PATSH framework is developed with the .Net cross platform technology. The main data structure is the **Sound object**, which contains a segment of the multidimensional audio signal whose interpretation is continuously refined during the processing pipeline. PATSH deals with the distribution of the data among the several plugins that perform the processing to interpret the audio events. The execution can be done, in parallel, synchronously or asynchronously, depending on the settings stored in a simple configuration file. In SWEET-HOME, the plugins were actually developed in C or C++ and PATSH includes the mechanism to transfer sound events from the plugins to the PATSH framework and vice-versa.

In the SWEET-HOME configuration, PATSH runs plugins that perform the following tasks:

1. Multichannel data Acquisition through the NI-DAQ6220E card. Seven channels are acquired at 16kHz (16 bits quantification);
2. Sound Detection and Extraction, detecting the start and end of sound events on each channel in parallel;
3. Sound/Speech Discrimination, discriminating speech from other sounds to extract voice commands;
4. Sound Classification, recognizing daily living sounds (not developed in this paper, see [13] for details);
5. Automatic Speech Recognition (ASR), applying speech recognition to events classified as speech and extracting vocal orders; and
6. Presentation, communicating the sound event to the Intelligent Controller. If a vocal order is detected and according to the context (activity and localisation of the user in the flat), a home automation command is generated to make the light up, close the curtains or emit a warning message thanks to a voice synthesizer.

The PATSH framework was developed to process on-line sound objects continuously detected on the 7 audio channels. However, it exists a bottleneck between the acquisition task and the event processing task. Given that one sound event can be simultaneously detected by several channels, the amount of the sound events in the queue can quickly rise.

2.2. Sound Event Detection

The detection of the occurrence of an audio event is based on the change of energy level of the 3 highest frequency coefficients of the Discrete Wavelet Transform (DWT) in a sliding window frame (last 2048 samples without overlapping). Each time the energy on a channel goes beyond a self-adaptive threshold, an audio event is detected until the energy decrease below this level for at least an imposed duration [2]. At the end of the detection, the Signal to Noise Ratio (SNR) is computed by dividing the energy in the event interval and the previous energy in a window outside this interval. This process is operated on each channel independently.

2.3. Sound/Speech Discrimination

Once sound occurrences are detected, the most important task is to distinguish speech from other sounds. In everyday life, there is a large number of different sounds, modelling all of them is irrelevant. For the SWEET-HOME project, distant voice command and distress situation detection, speech is the most important sound class. The method used for speech/sound discrimination is a GMM (Gaussian Mixture Models) classification.

The Sound/Speech Discrimination stage has a very important role: firstly, vocal orders must not be missed, secondly, daily living sounds must not be sent to the ASR because undesirable sentences could be recognized. To recognize only vocal orders and not all sentences uttered in the flat, all sound events shorter than 150 ms and longer than 2.2 seconds were ignored as well as those whose SNR is below 0 dB. These values were chosen after a statistical study on our data bases.

2.4. Voice order recognition

In a Smart Home, the microphones are generally set in the ceiling and on the wall. This places the study in a distant-speech context where microphones may be far from the speaker and may record different noise sources. Moreover, the application calls for quick decoding so that voice commands are sent as soon as possible to the intelligent controller. This is why we used the Speeral tool-kit [10] developed by the LIA

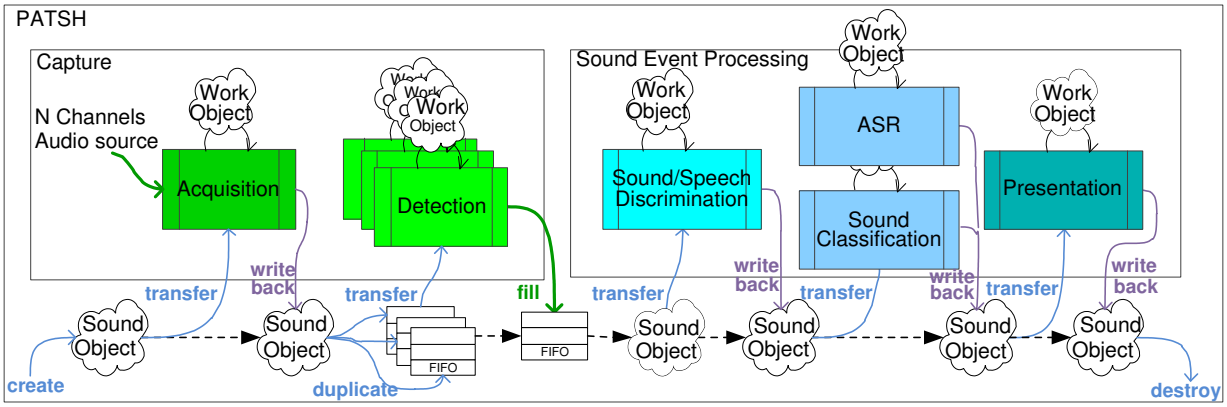

Figure 1: The PATSH architecture.

(Laboratoire d'Informatique d'Avignon). Indeed, its 1xRT configuration allows a decoding time similar to the signal duration. Speeral relies on an A^* decoder with HMM-based context-dependent acoustic models and trigram language models. HMMs are classical three-state left-right models and state tying is achieved by using decision trees. Acoustic vectors are composed of 12 PLP (Perceptual Linear Predictive) coefficients, the energy, and the first and second order derivatives of these 13 parameters.

The acoustic models of the ASR system were trained on about 80 hours of annotated speech. Furthermore, acoustic models were adapted to the speech of 23 speakers recorded in the same flat during previous experiments by using Maximum Likelihood Linear Regression (MLLR) [8]. A 3-gram Language Model (LM) with a 10K lexicon was used. It results from the interpolation of a *generic* LM (weight 10%) and a *domain* LM (weight 90%). The *generic* LM was estimated on about 1000M of words from the French newspapers *Le Monde* and *Gigaword*. The *domain* LM was trained on the sentences generated using the grammar of the application (see Fig. 3). The LM combination biases the decoding towards the *domain* LM but still allows decoding of out-of-domain sentences. A probabilistic model was preferred over using strictly the grammar because it makes it possible to use uncertain hypotheses in a fusion process for more robustness.

3. Experiments in real conditions

3.1. Experimental flat

Experiments were run in the DOMUS smart home. Figure 2 shows the details of the flat. It is a thirty square meters suite flat including a bathroom, a kitchen, a bedroom and a study, all equipped with 150 (konnex) KNX sensors and actuators. The flat has been equipped with 7 radio microphones set in the ceiling for audio analysis. A specialized communication device, *e-lio*, from the *Technosens* company was used to initiate a communication between the user and a relative.

3.2. Voice orders

Possible voice orders were defined using a very simple grammar as shown on Figure 3. Each order belongs to one of three categories: initiate command, stop command and emergency call. Except for the emergency call, every command starts with a unique key-word that permits to know whether the person is

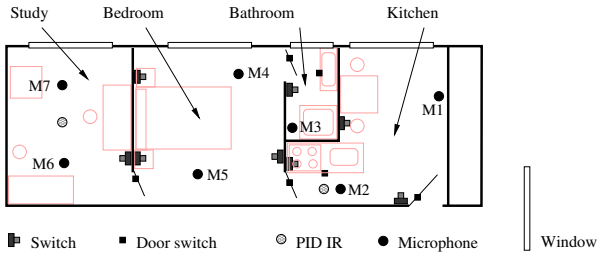

Figure 2: Position of the microphones and other sensors inside the DOMUS smart home.

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. `Nestor au secours`)

The grammar was built after a user study that showed that targeted users would prefer precise short sentences over more natural long sentences [11]. In this study, although most of the older people spontaneously controlled the home by uttering sentences, the majority said they wanted to control the home using keywords. They believe that this mode of interaction would be the quickest and the most efficient. This study also showed that they also had tendency to prefer or to accept the 'tu' form (informal in French) to communicate with the system given this system would be their property.

3.3. Scenarios and experiments

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 *e-lio*. 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. For instance, for the decision regarding the activation of the light,

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

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

given that the bedroom had two lights (the ceiling and the bedside one) as well as the kitchen (above the dinning table and above the sink), the four following situations were planed:

1. **Situation 1**. The person is having a meal on the kitchen table. The most appropriate light is the one above the table.
2. **Situation 2**. The person is cleaning up the bedroom. The most appropriate light is the ceiling one.
3. **Situation 3**. The person is cleaning the sink and doing the dishes. The most appropriate light is the one above the sink.
4. **Situation 4**. The person has just finished a nap. The most appropriate light is the bedside one.

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 case of, a wizard of Oz was used in case of persistent problem.

Sixteen participants (including 7 women) without special needs were asked to perform the scenarios without condition on the duration. A visit, before the experiment, was organized to ensure that the participants will find all the items necessary to perform the scenarios. It was necessary to explain the right way to utter vocal orders and to use the *e-lio* system. 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.

3.4. Acquired Corpus

During the experiment, audio data were recorded and saved in two ways. Firstly, the 7-channel raw audio signal was stored for each participant to make subsequent analysis possible. In total, 8h 52min 36s of data was recorded for the 16 participants. Secondly, the individual sound events automatically detected by PATSH were recorded to study the performances of this framework.

Apart from daily living sounds and sentences uttered in the flat by the participant, PATSH also detected the system messages (vocal synthesizer) and the *e-lio* communications. Overall, 4595 audio events were detected whose 993 were speech and 3503 were other noise occurrences. The number of events corresponding to each category –speech or everyday living sound– is displayed Table 1.

Table 1: *Number of audio events (speech and sound).*

Speaker ID	Speech and sound	Sound	Speech	Mis classified speech	Mis classified sound
S01	213	184	29	8	1
S02	285	212	73	10	6
S03	211	150	61	8	6
S04	302	211	91	10	11
S05	247	100	48	11	4
S06	234	189	45	17	6
S07	289	216	72	21	6
S08	249	190	59	25	3
S09	374	283	91	19	7
S10	216	163	53	10	4
S11	211	155	56	18	2
S12	401	346	55	13	13
S13	225	184	41	4	7
S14	235	173	62	9	10
S15	641	531	111	39	17
S16	262	216	46	10	5
ALL	4595	3503	993	232	108

In this study, we are only interested in recognizing vocal orders or distress sentences. All other spontaneous sentences and system messages are not irrelevant. Therefore, the global audio records were annotated using Transcriber in order to extract the syntactically correct vocal orders, results are shown in Table 2. The average SNR and duration are 15.8dB and 1s, this SNR value is low compared to studio conditions (SNR$\geq$35dB). As the home automation system needs only one correct sentence to interact, only the less noisy channel was kept. The number of vocal orders is different for each speaker because if a vocal order was not correctly recognized, the requested action was not operated by the intelligent controller (light on or off, curtains up or down...) and thus the speaker often uttered the order two or three times. Thanks to this annotation, an oracle corpus was extracted. The comparison between experimental real-time results with thus obtained with the same ASR on the oracle corpus will allow to analyse the performance of the PATSH system.

Table 2: *Number of syntactically correct vocal orders*

Speaker ID	Number	SNR (dB)	Speaker ID	Number	SNR (dB)
S01	20	17	S02	32	17
S03	22	19	S04	26	18
S05	26	12	S06	24	15
S07	19	25	S08	33	12
S09	40	20	S10	40	11
S11	37	14	S12	26	17
S13	21	14	S14	27	12
S15	28	14	S16	22	14
All	443	15.8			

Table 3: *Home automation order error rate (DER)*

Speaker ID	Expe. (%)	Oracle (%)	Speaker ID	Expe. (%)	Oracle (%)
S01	35	20	S02	12.5	6.2
S03	22.7	22.7	S04	23	7.7
S05	15	3.8	S06	21	8.3
S07	79	52.6	S08	30	33.3
S09	40	22.5	S10	67	47.5
S11	46	27	S12	21	7.7
S13	43	19	S14	48	29.6
S15	71	55.5	S16	18	13.6
Average	38%	23.9%			

4. Results

4.1. Discrimination between speech and sounds

The detection part of the system is not specifically evaluated because of the lack of time to label all the sound events on the 7 channels. However, all the results presented take into account the performances of the detection because the signals are extracted automatically by the system. The sound/speech discrimination misclassified 108 sound and 232 speech occurrences which gives a total error rate of about 7.4% which is in line with other results of the literature [13]. 23.4% of speech occurrences were classified as sound. Theses poor performances are explained by the fact that PATSH was not successful in selecting the best audio event among the set of simultaneous events and thus the events with low SNR introduced errors and were not properly discriminated. For the sounds, 3.1% of sound occurencies are classified as speech. Sounds such as dishes, water flow or electric motor were often confused with speech. For instance, when certain persons stirred the coffee and chocked the spoon on the cup or when they chocked plates and cutlery, the emitted sounds had resonant frequencies very close to the speech one. This is emphasizing the difficulty of the task and models must be improved to handle these problematic samples.

4.2. Home automation order recognition

The global performance of the system is directly related to vocal order recognition. The DER (Domotic Error Rate) is shown in Table 3, the 2^{nd} and 5^{th} columns "Expe." indicates the results for the real-time experiment. This error rate is evaluated after filtering at the input of the intelligent controller and includes the global effects of all stages: detection, discrimination between speech and sound, ASR. When the uttered voice orders were not respecting the grammar, for example when a sentence such as “Nestor heure” is uttered instead of the command “Nestor donne l’heure”, these utterances were discarded. Moreover, some speakers’ utterances exceeded the 2.2s duration threshold because of their hesitation, therefore corresponding vocal orders were not analysed and considered as missed. In case of music in the room, vocal orders were often longer because of the mixing between speech and music. Consequently future experiments will need to set the threshold to 2.5s and to include a short training step to allow the participant to become familiar with this technology.

The ASR system used generic acoustic models without adaptation to the speaker and then regional or foreigner accent may have an influence: it’s in particular the case for S10 (Arabic) and S14 (Alsatian). The participants S07 and S15 show low performance because they were not able to follow the given instructions, the presence of large part of silence mixed with noise between the words is analysed as phoneme and therefore increases the error rate.

Part of the errors was due to the way PATSH managed simultaneous detections of one sound event. At this time of the process, the SNR is not known with a sufficient precision and the choice is not perfect. Then, in some cases, a part of the speech signal is missed (beginning or end of the order) and this introduces a bad recognition. Moreover, very often the detection is not perfectly simultaneous and more than one channel is analysed by the ASR. Therefore, some improvement were introduced in PATSH for future experiments: that consisted in making the decision after the end of detection on the 7 threads (each thread corresponding to one channel) thanks to a filtering window of 500ms. The disadvantage is that the system is slowed down with a delay of 500ms but this will avoid the recognition of bad extracted sentences and this is compensated by the analysis of only the signal of the best channel.

An important aspect is the decoding time because the device must be activated with a delay as short as possible. In this experiment, the decoding times reached up to 4 seconds which was a clear obstacle for usage in real condition. Hopeful, this has been reduced.

5. Preliminary Results from experiments with the aged and visually impaired population

The method has also been applied in the same context but with aged and visually impaired people. The aim was both to validate the technology with this specific population and to perform a user study to assess the adequacy of this technology with the targeted users and to compare with the other user studies of the literature [14, 11].

Between the two experiments, several corrections were applied to PATSH so that the sound/speech discrimination was greatly improved as well as the speech decoding time. The measured decoding time was 1.47 times the sentence duration; as the average duration of a vocal order was 1.048s, the delay between the end of the utterance and the execution of the order was 1.55s. This is still not a satisfactory delay but this does not prevent usage in real conditions.

5.1. Experimental set up

In this experiment, eleven participants either aged (6 women) or visually impaired (2 women, 3 men) were recruited. The average age was 72 years (49-91, min-max). The aged persons were

fully autonomous but were living alone. The participants were asked to perform 4 scenarios involving daily living activities and distress or risky situations.

1. The participant is eating her breakfast and is preparing to go out. She asked the house to turn on the light, close the blinds or ask for the temperature while doing these activities.
2. The participant is coming back from shopping and is going to have a nap. She asked the same kind of commands but in this case, a warning situation alerts about the front door not being locked.
3. 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.
4. The participant is waiting in the study for friends going to visit her. She tests various voice orders with the radio, lights and blinds.

During this experiment, 4 hours and 39 minutes of data was collected including the same sensors as the one previously described in Section 3.4.

5.2. First feedbacks

All the participants went through a questionnaire and a debriefing after the experiment. We are still in the process of analysing the results but overall, none of the aged or visually impaired persons had any difficulty in performing the experiment. They all appreciated to control the house by voice.

It is worth emphasizing that aged people preferred the manual interaction because this was quicker. However, they liked the voice warning in case of risky situations. Regarding the visually impaired participants, they found that the voice command would be more adequate if it could enable performing more complex or dangerous tasks than controlling blinds or radio. For instance, by enabling them to use the household appliances. Overall, half of the participants found the system adapted for their use.

6. Discussion

Overall, the performance of the system was still low but the results showed there is room for improvement. Sound/Speech discrimination has been improved since the beginning of the experiment and continue to be improved. The biggest problems were the response time which was unsatisfactory (for 6 participants out of 16) and the mis-understanding of the system which implied to repeat the order (8/16). These technical limitations were reduced when we improved the ASR memory management and reduced the search space. After this improvement, only one participant with special needs complained about the response time. None of the encountered problem challenged the PATSH architecture. That is why we are studying the possibility of releasing the code publicly.

The grammar was not the focus of the project but it has been built to be easily adaptable at the word level (for instance, if someone wants to change "Nestor" for another word). All the 16 participants found the grammar easy to learn. Only four of them found the keyword "Nestor" unnatural while the others found it natural and funny. However, this approach suffers a lack of natural adaptivity to the user's preferences, capacities and culture as any change would require technical intervention. For instance, [15] emphasized that elder Germans tend to utter longer and politer commands than their fellow countrymen which contrast with our findings. Despite longitudinal studies are require to understand human preferences regarding voice orders, methods to adapt on-line the grammar to the user must be developed.

The acquired corpus made it possible to evaluate the performance of the audio analysis software. But interest goes far beyond this experiment because it constitutes a precious resource for future work. Indeed, one of the main problems that impede researches in this domain is the need for a large amount of annotated data (for analysis, machine learning and reference for comparison). The acquisition of such datasets is highly expensive both in terms of material and of human resources. For instance, in a previous experiment involving 21 participants in the DOMUS smart home, the acquisition and the annotation of a 33-hour corpus has cost approximatively 70k€. Thus, making these datasets available to the research community is highly desirable. This is why we are studying the possibility of making part of it available to the society as we did in our previous project [16].

7. Conclusion

This paper presents the PATSH system, the audio processing module of the voice controlled SWEET-HOME system which performs real-time identification of voice commands in the home for assisted living. In this system, the identified sound events are sent to an intelligent controller for final context-aware decision about the action to make on the house [17]. The experiments made in the Smart Home to evaluate the system showed promising results and validate the approach. This technology can benefit both the disabled and the elderly population that have difficulties in moving or seeing and want security reassurance.

Our application of this technology within a realistic Smart Home, showed that one of the most sensible tasks is the speech/noise discrimination [9]. According to the SNR level, the performance can be quite poor, which has side effects on both the ASR and the sound classification (and then on the decision making). Another issue is linked to the lack of handling of simultaneous sound event records. These fill the sound object queue, which is the system bottleneck, and thus slow down the processing while real-time performances are required. To increase the performance and free this bottleneck, we had implemented a filtering strategy to remove low SNR audio events as well as too delayed events. The preliminary results showed a significant increase in performance. In a second step, PATSH will be modified to allow in real-time a multisource ASR thanks to the Driven Decoding Algorithm [18].

Although the participants had to repeat, sometimes up to three times, the voice command, they were overall very excited about commanding their own home by voice. We are still in the process of analysing the results of the experiment which included seniors and visually impaired people to get essential feedback from this targeted population. Future work will include improvements of the speech recognition in noisy environment and customisation of the grammar as well as experiments using specialised communication devices to enhance user's communication capacity.

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 B. Meillon, N. Bonnefond, D. Guerin, C. Fontaine and S. Pons for their support.

9. References

[1] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes- present state and future challenges," *Computer Methods and Programs in Biomedicine*, vol. 91, no. 1, pp. 55–81, 2008.

[2] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat, "Information extraction from sound for medical telemonitoring," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, pp. 264–274, April 2006.

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

[4] M. Popescu, Y. Li, M. Skubic, and M. Rantz, "An acoustic fall detector system that uses sound height information to reduce the false alarm rate," in *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, 20–25 Aug. 2008, pp. 4628–4631.

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

[6] M. Hamill, V. Young, J. Boger, and A. Mihailidis, "Development of an automated speech recognition interface for personal emergency response systems," *Journal of NeuroEngineering and Rehabilitation*, vol. 6, no. 1, p. 26, 2009.

[7] G. Filho and T. Moir, "From science fiction to science fact: a smart-house interface using speech technology and a photorealistic avatar," *International Journal of Computer Applications in Technology*, vol. 39, no. 8, pp. 32–39, 2010.

[8] B. Lecouteux, M. Vacher, and F. Portet, "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions," in *Interspeech 2011*, Florence, Italy, 2011, pp. 2273–2276.

[9] M. Vacher, F. Portet, A. Fleury, and N. Noury, "Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges," *International Journal of E-Health and Medical Communications*, vol. 2, no. 1, pp. 35–54, 2011.

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

[11] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[12] P. Chahuara, F. Portet, and M. Vacher, "Context aware decision system in a smart home : knowledge representation and decision making using uncertain contextual information," in *The 4th International Workshop on Acquisition, Representation and Reasoning with Contextualized Knowledge (ARCOE-12)*, Montpellier, France, 2012, pp. 52–64.

[13] M. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, and J. Boudy, "Sound Environnment Analysis in Smart Home," in *Ambient Intelligence*, Pisa, Italy, 2012, pp. 208–223.

[14] R. López-Cózar and Z. Callejas, "Multimodal dialogue for ambient intelligence and smart environments," in *Handbook of Ambient Intelligence and Smart Environments*, H. Nakashima, H. Aghajan, and J. C. Augusto, Eds. Springer US, 2010, pp. 559–579.

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

[16] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A french corpus of audio and multimodal interactions in a health smart home," *Journal on Multimodal User Interfaces*, vol. 7, no. 1, pp. 93–109, 2013.

[17] 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, "The SWEET-HOME project: Audio processing and decision making in smart home to improve well-being and reliance," in *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'13)*, 2013.

[18] M. Vacher, B. Lecouteux, and F. Portet, "Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment," in *EUSIPCO (European Signal Processing Conference)*, Bucarest, Romania, August 27-31 2012, pp. 1663–1667.