Context-Aware Voice-based Interaction in Smart Home -VocADom@A4H Corpus Collection and Empirical Assessment of its Usefulness

François Portet, Sybille Caffiau, Fabien Ringeval, Michel Vacher, Nicolas Bonnefond, Solange Rossato, Benjamin Lecouteux, Thierry Desot

Context-Aware Voice-based Interaction in Smart Home – VocADom@A4H Corpus Collection and Empirical Assessment of its Usefulness

François Portet and Sybille Caffiau and Fabien Ringeval and Michel Vacher
and Nicolas Bonnefond and Solange Rossato and Benjamin Lecouteux and Thierry Desot
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
Email: Firstname.Lastname@imag.fr

***Abstract*—Smart homes aim at enhancing the quality of life of people at home by the use of home automation systems and Ambient Intelligence. Most of these smart homes provide enhanced interaction by relying on context-aware systems learned on data. Whereas voice-based interaction is the current emerging trend, most available corpora are either concerned only with home automation sensors or only with audio technology, which limits the development of context-aware voice-based systems. This paper presents the VocADom@A4H corpus, which is a dataset composed of users' interactions recorded in a fully equipped Smart Home. About 12 hours of multichannel distant speech signal synchronized with logs of an openHAB home automation system were collected from 11 participants who performed activities of daily living with the presence of real-life noises, such as other persons speaking, use of vacuum cleaner, TV, etc. This corpus can serve as a valuable material for studies in pervasive intelligence, such as human tracking, human activity recognition, context aware interaction, and robust distant speech processing in the home. Experiments performed on multichannel speech and home automation sensors data for robust voice activity detection and multiresident localization show the potential of the corpus to support the development of context-aware smart home systems.**

I. Introduction

The goal of Ambient Assisted Living (AAL) is to foster the emergence of ICT-based solutions to enhance the quality of life of people at home, at work and in the community.

Smart spaces, such as Smart Homes, are emerging as a way to fulfill this goal [1]. Furthermore Smart Homes can support older adults and people with disabilities by monitoring the person's health and behavior through sensors and increase autonomy by enabling natural control of the people's environment through home automation [2]. In this context, audio-based technology has a great potential to become one of the major interaction modalities in Smart Home, as it fulfills Weiser's vision in which computing technology becomes so seamlessly integrated into our environment that we use it without noticing it [3]. Audio technology has indeed many properties that fit this vision: it is physically intangible and its large sensing range 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 procedures or jargon. Despite all this and the rising number of smart speakers on the market, a small number of smart home projects have seriously considered audio technology in their design. Consequently, few corpora related to voice-based home interaction which include several dwellers are available in the community to develop such context-aware smart home systems.

This paper summarizes the effort developed in the VocADom project[1] to collect a corpus of users' interaction in a smart home to be made available to the community. The VocADom@A4H corpus collection considered typical tasks undertaken as part of Smart Home systems, such as automatic location of dwellers, Human Activity Recognition (HAR), Voice Activity Detection (VAD), speaker identification, Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), context-aware decision, etc. This paper describes the complete protocol, collection and annotation and emphasizes its importance for the pervasive community. The contribution of this work is multifold and lies in: (i) the exposed methodology to acquire a controlled but natural as possible corpus, (ii) the release of a multisensor (including speech and home automation traces) and multiuser corpus useful for several tasks related to AAL, and (iii) experiments showing the usefulness of the corpus.

The remainder of this paper is structured as follows: we perform a critical review of the available corpora for AAL in Section II, introduce the Smart Home environment in Section III, the experimental protocol used for data collection in Section IV, and the experiments performed on the VocADom corpus in Section VI, before concluding in Section VII.

II. State of the Art: Available Corpora

The development of AI based smart home technologies requires datasets obtained from experimental platforms providing the necessary conditions to represent real situations. Collecting these datasets is very demanding in effort, resources, and time, and asks 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 research projects – either national research bodies or companies – that incidentally biases the collection towards

[1] https://vocadom.imag.fr/

the project objectives, e. g., validation of a specific interface. Moreover, the experimental material is often available to a small part of the research community. As a result, this slows down the research advances in the domain. To overcome this situation, a number of corpora were collected and released to the community each with its own advantages and limitations.

For instance, in the MavHome project [4] the objectives were to use data mining techniques to control the environment according to the user's activities. Two datasets were obtained from an office-like environment and an apartment test-bed. The former corpus was collected during two months where six students worked in the office, whereas the latter concerned a single person who lived there for a two-month period. Both corpora contain sensor readings about users' motion, light, temperature, humidity and doors state. In the University of Amsterdam, Kasteren *et al.* [5] collected a corpus using 14 state-change sensors installed in an apartment where a 26-years-old man lived alone during 24 days. The sensors were set on objects easily associated with Activities of Daily Living (ADLs), such as the microwave, the fridge, and the toilet flush. The ADL annotations were made by the user himself who used a headset along with a speech recognition module to record the beginning and end of each activity and served as the boundaries of activities in the annotated data. This put into question the naturalness of the dataset since the user could not forget at any stage he was not in an ecological situation.

To move from single resident to multi-resident, the well known CASAS project [6] acquired an impressive amount of corpora in flats in different locations. The research focused mainly in the recognition and analysis of ADLs. 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 a flat for several months. With regard to the employed sensors, they fall in the categories of motion, contact doors, temperature, electricity and water consumption. Many other corpora of the kind were collected in different houses during several days and with sometimes several residents, e.g., GER'HOME [7], Orange4Home [8], Transfer Learning dataset [9] or the ARAS datasets [10].

One common aspect of the above described corpora is that they do not always include video or audio sensors along with automation logs during the experiments, which makes them difficult to analyze and to annotate in regard to some other aspects of the users' behavior than the ones they were designed for. Hence, none of the previously collected corpora can be used to developed context-aware voice based interactive systems since they do not include audio recording or decision making.

In the meantime, in the speech community, distant speech processing in Smart Homes has drawn an increasing interest thanks to recent advances in speech processing from microphone arrays. Therefore, several speech corpora were recorded. For instance, the CHiME corpora [11] are made of English speech recordings in different noise conditions. In particular, the CHiME-5 data set is composed of recordings of 4-person dinner parties (host couple and guests). Data were recorded from fixed microphone arrays, head worn microphones and cameras. The ITAAL Italian speech corpus [12] was made of records at home with a headset microphone and an array of four microphones. Sentences were home automation commands (e. g., "close the door"), distress calls (e. g., "help!", "ambulance!") and phonetically rich sentences. The DIRHA English corpora [13] were designed to provide multichannel acoustic data that could be used to investigate speaker localization, acoustic echo cancellation, speech enhancement and ASR. Recordings were made with microphone arrays in each room of a flat. Twelve UK native speakers and twelve US native speakers were recorded. For each of them, the corresponding recorded material includes read and spontaneous home automation commands, keywords, phonetically-rich sentences and about 10 min of conversational speech. The voiceHome [14] corpus was recorded for distant speech processing analysis in home. It includes live speech from native French speakers in reverberant and noisy conditions. Another example is the SINS corpus [15] of the D-CASE challenge, consisting of continuous audio recordings of one person living in a home over a week.

Each of these corpora was recorded in home conditions and can be used for studies involving distant speech recognition and noise cancellation due to the use of microphone arrays. However, no home automation sensor traces were recorded, and only CHiME-5 and the SINS corpora were recorded in ecological situations, i. e., while performing ADL or involving the use of household appliances such as the vacuum cleaner. It is thus not possible to build models jointly using acoustic and home automation sensors data.

In fact, very few corpora include home automation traces and audio channel. The MIT's Place Lab corpora [16] contain video and audio files of in-home experiments but the audio channel was used for annotation purpose only. The HIS corpus [17] was one of the first data collection including speech, everyday life sounds and home automation sensor traces but its purpose was single inhabitant monitoring. Hence, it cannot be used to develop interactive systems. The Sweet-Home corpus [18] was recorded by participants enacting ADL in a smart home equipped with home automation sensors and actuators. Continuous speech was mainly composed of voice commands. It is the only corpus we are aware of which contains home-automation sensors, audio signals and decision making. However it was only recorded in a single user settings and with a very low amount of everyday noise.

This short state-of-the-art shows the necessity to record a new corpus useful for the development of context-aware voice-based smart home systems.

III. Smart Home Environment

In this work, the pilot smart home was Amiqual4Home[2] [19]. This 87 m^2 Smart Home is equipped with home automation systems, multimedia controller, environmental meters and actuators, and means for observing human activity. The kitchen

[2]https://amiqual4home.inria.fr

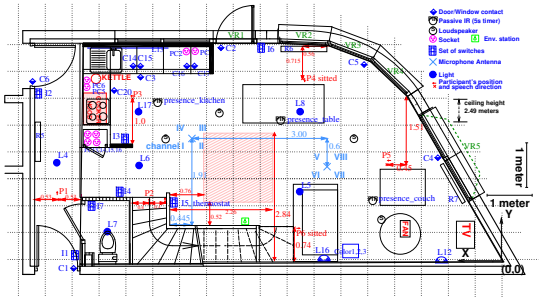

Fig. 1. Ground floor: kitchen and living room.

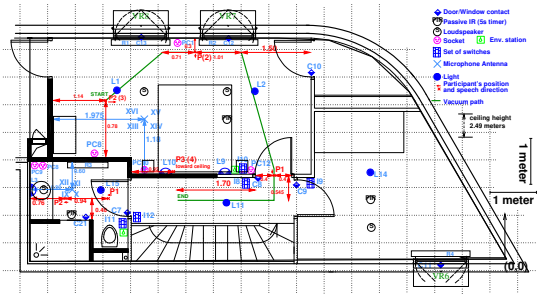

Fig. 3. Upper floor: bedroom and bathroom.

and the living room are on the ground floor (Fig. 1 and 2), the bedroom and the bathroom on the upper floor (Fig. 3). A dedicated hidden control room permits to centralize the recording of the sensors and to control the devices remotely.

A. Home Automation Sensors

This Smart Home is fully functional and equipped with more that 500 controllable or observable items. Home automation sensors and actuators, e. g., lighting, shutters, security systems, energy management, heating, etc., are connected by a KNX[3] bus system (standard ISO/IEC 14543). Besides KNX, several field buses coexist, such as UPnP (Universal Plug and Play) for the multimedia distribution, X2D for the contact detection (doors, windows and cupboards), RFID for the interaction with tangible objects (not used in the VOCADOM project). The management of the home automation network, sending commands to the different actuators and receiving changes of sensor values, is operated through openHAB[4]. This layer guarantees the interoperability of the data coming from the different field buses and allows the communication between them and towards virtual applications, such as activity tracking. Thanks to this gateway, all devices including multimedia elements can be controlled remotely. In addition, 9 cameras are set up in the ceiling of the rooms.

B. Audio Recording

Four 4-channel microphone arrays as shown Figure 4, are set in the ceiling of the kitchen, the living room, the bathroom and the bedroom and directed towards the ground. Each array is made of 4 LC 97 TWS Lavalier microphones. In addition, the participant wore a HSP 4 Sennheiser microphone in front of the mouth to ease the speech transcription.

[3]https://www.knx.org/
[4]https://www.openhab.org/

Fig. 2. Instrumented kitchen showing the landmark #3 on the ground facing the cooker hood.

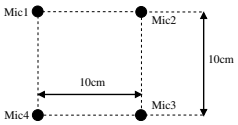

Fig. 4. Microphone array.

IV. Experimental Protocol

As sketched in the introduction, the objective of the data collection is to support the development and evaluation of new context-aware systems. The challenge in such a task is to conduct the recording so that the collected data is not only useful for different processing tasks but also representative enough of a variety of realistic domestic situations.

The protocol was designed to collect data for seven tasks: Human Localization (HL), Human Activity Recognition (HAR), Speech Enhancement (SE), Voice Activity Detection (VAD), Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), and Automatic Decision Making (ADM). As shown in Table I, each task can rely on different combinations of multimodal data processing, such as voice commands (extracted from speech), noises (noise of domestic devices) and home automation sensors.

To represent domestic situations, three recording conditions (in three phases) were designed in which participants – enacting the role of dweller or visitor – used voice commands to communicate with the home and to perform ADLs. The three phases enabled different degree of spontaneity in the participants' behavior and contained several noisy conditions (human speaking, vacuum cleaner, TV, fan and shower). Hence, for machine learning approaches, data of last phases can be used for learning (most controlled and numerous) whereas data of first phases are more suited for testing (more spontaneous and lower amount of data).

A. Definition and Collection of Voice Commands

The voice command were designed to always be composed of a keyword used to make clear the speaker was addressing the Smart Home. Keywords are common trigger of voice based applications and must be chosen with care. Most keywords

TABLE I
USEFUL DATA FOR SEVEN AAL RELATED DATA PROCESSING TASKS

	Voice command	Speech	Noise	Home Automation Sensors
HL		X	X	X
HAR		X	X	X
SE		X	X	
VAD		X	X	X
ASR	X	X		
NLU	X			X
ADM	X		X	X

are of at least 3/4 syllables long to enable sufficient duration for correct detection. Furthermore, since Smart Homes are of particular interest for supporting older adults staying at home as long as possible, the phonetic components of the keyword must be particularly easy to utter for an older adult. Five consonants and five vowels were chosen from a recent study [20], where different phones in French were studied with respect to the performance of an ASR system. Results showed that, in the International Phonetic Alphabet (IPA), the consonants s, ʃ, m, ʁ and l and the vowels i, y, u, ɛ and e were the best recognized. Words respecting these constrains were then extracted automatically from a dictionary and further filtered and discussed to provide the following list: *téraphim, ulysse, ichefix, chanticou, vocadom, écirrus, hé cirrus, allo cirrus, allo messire, dis vesta, dis hestia, dis bérénio, dis téraphim, dis vocadom, minouche*.

The keyword was followed by a sentence specifying the voice command, which can be divided into four main categories: (i) `contact` which allows a user to place a call, e. g., *"Call my son"*, (ii) `set` to set the state of an object in the smart-home, e. g., *"Turn on the TV"*, (iii) `get` to query the state of objects as well as properties of the world at large, e. g., *"What time is it?"*, and (iv) `check` to check the state of an object, e. g., *"Is the door closed?"*. All these commands can have possible adjuncts of time, space, or other adjuncts specifying the query. Only the category `set` and `check` were used in this experiment.

The main feedback of the Smart Home was the actuation of the meant actuators in the voice command – if the user wanted a light on, the lighting of the light was a sufficient feedback in most cases. Four other types of feedback were defined as short non-linguistic sounds: (1) Incorrect/impossible command, (2) Similar to a command but forgotten keyword, (3) Affirmative answer, and (4) Negative answer. The first two sounds were used in case of an error, whereas the last two sounds were used for answering closed questions (`check`).

B. Strategy to Ensure Spontaneous Interaction

Since voice controlled Smart Home is still in its infancy, there is not a clear understanding about what an ecological voice controlled Smart Home dataset should look like. Thus, to add credence to the experiment, we included the following properties to the recording to make it as realistic and as spontaneous as possible:

- Wizard of Oz: The home automation system was activated from the technical room by the experimenters serving as wizards. The participant was not informed whether or not the system was indeed automatic till the end of the experiment. Since video and audio streams were captured in real time, each time a wizard heard a voice command, he performed the corresponding action, otherwise the appropriate error feedback was played.
- Abstract voice command description: In order to make the participants use their own words to command the house, the experiment started with abstract non-linguistic descriptions of the voice commands to utter.
- Activities of Daily Living: All participants were asked to perform high level activities (making coffee, watching TV) while uttering commands so that speech and Smart Home data were collected in daily living situations.
- Multi-speaker: In a specific phase, an experimenter was present in the home so that unscripted discussions and activities were also recorded.

C. Various Domestic Situations

In order to collect as ecological as possible data from domestic life, participants were asked to consider the apartment as their home and to play domestic activities such as preparing a cup of tea, receiving visit, etc. When an activity necessitated the use of smart devices, the participant uttered a voice command and the wizard(s) activated the corresponding actions. Three recording phases were defined as:

- Phase 1 – *Graphical based instruction to elicit spontaneous voice commands*. In order to avoid any influence from the experimenter, the participants were left without any lexical cues on how to control the Smart Home and were asked instead to experiment themselves using the system to elicit their own lexicon, while having graphical representations of the task to be performed at hand as guidance.
- Phase 2 – *Two-inhabitant scenario enacting a visit by a friend*. Activities and voice commands were planned but participants were free to perform the activities and utter sentences the way they wanted.
- Phase 3 – *Voice commands in noisy domestic environment*. The participants had to read a list of voice commands at different places in the apartment; six places on the ground floor living room / kitchen, two in the bathroom and three in the room. There were, in each case, five sentences to read in each position and each noise context.

In the first phase, only illustrations depicting the activity to be performed were provided to the participants whom used their own vocabulary to make the home perform what they wanted. In phase 2, the activities were written down in a scenario on a sheet of paper, e. g., "go in the bedroom and listen to the radio", "make sure the blinds are closed before leaving". However no strict grammar was provided since users tend to deviate from predefined sets of voice commands [21], [22], [2]. Finally in Phase 3, a large number of voice commands were intended to be uttered to study speech recognition in noisy environment using a fixed grammar similar to [18].

TABLE II
ROOMS, LANDMARKS AND NOISE CONDITION OF PHASE 3

#	living room/kitchen	#	Bathroom
1	No noise	1	No noise
2	Radio	2	Shower
3	Fan	3	Tap
4	Someone reading	#	Bedroom
	- while walking within a	1	No noise
	- square on the ground	2	Blinds
5	Television	3	Radio
6	Television + cooker hood	4	Vacuum (mobile)

We used a large variety of domestic noise which are listed in Table II with the corresponding room and landmark at which the participant had to be. For the living room/kitchen level (cf. Figure 1), the TV, fan, cooker hood were used as well as a person reading in the red square dashed area visible on the floor. The fan was always positioned on a coffee table in front of the TV. The radio was alternatively played from two pairs of speakers either in the kitchen or in the living room. The place of the origin of the noise never changed during the recording except the fan which was rotating. On the upper floor (cf. Figure 3), the shower and tap of the sink were used as noise sources in the bathroom. Blinds in the bedroom were also controlled to generate noise. Finally, the participant was asked to use the vacuum along a defined route placed as show the dark green lines around the bed on Figure 3.

V. COLLECTED CORPUS

Before starting the dataset collection and playing domestic activities, the participant signed a consent form[5], and an experimenter explained the objectives and the role of the participants, before visiting the apartment (alone) and choosing a keyword to control the home.

The experiment was run during June 2017 in the Amiqual4Home Smart Home. More than 12 hours of data were collected. Eleven volunteer members of the lab with no relation with the project were recruited whose details are shown in Table III. The average age was of 23-25 years old with a good balance in gender (4 females) and a mean recording time of 1 hour and 8 minutes per session. The table III also shows the diversity of chosen keywords.

The dataset is structured with different folders according to the recording session, e. g. for participant `S01`:

```
record/S01/
|-- mic_array/
|-- mic_headset/
|-- openhab_log/
|-- video/
```

The `mic_array` directory contains the 16-channel audio recordings of the microphones arrays placed as shown in Figures 1 and 3. Each of the 16 wav files contains a one-channel acquired at 44.1 kHz and encoded using Pulse Code Modulation (PCM) on 24 bits. All channels are perfectly synchronized. The `mic_headset` directory contains the recording of the one channel worn microphone (16 kHz, 16 bits PCM). The `openhab_log` directory contains two logs of the openHAB network in the .csv format: S01_change.csv.log (a timestamped item state each time there is a change, e. g., door from close to open), S01_wizard.csv.log (records all the commands sent by the wizards). The `video` directory contains the six video streams of the record; four in the ground floor, two in the bedroom, and none in the bathroom.

[5]The acquisition protocol was validated by the CNIL, which is the French institution protecting personal data and preserving individual liberties.

TABLE III
PARTICIPANTS RECORDING; DURATION IS GIVEN IN THE FORMAT: HOUR:MINUTES:SECONNDS.

Participant	Age group	Gender	Duration	Keyword
S00	20-23 years	M	01:03:54	vocadom
S01	20-23 years	M	00:48:53	vocadom
S02	20-23 years	M	01:12:26	hé cirrus
S03	20-23 years	M	01:11:52	ulysse
S04	23-25 years	F	01:04:46	téraphim
S05	<20 years	F	01:22:59	allo cirrus
S06	23-25 years	M	00:55:54	ulysse
S07	25-28 years	M	01:03:54	ichefix
S08	23-25 years	M	01:13:01	ulysse
S09	23-25 years	F	01:20:06	minouche
S10	23-25 years	F	01:11:03	hestia
All	(mean) 23-25 years	4 F / 7 M	12:28:45	8 keywords

Overall there are 16 GB of videos, 122 GB of audio recording of the arrays, 1.4 GB of audio recording of the worn microphone and 7.8 MB of logs. The logs are composed of 64,942 events of change and 1,075 wizard actions. The audio data of the 11 participants are composed of 110 voice commands for phase 1, 132 (two participants) in phase 2 and 3,190 in phase 3 (whose 2,493 are unique).

Location and the activity of the participants have been annotated using Simple-ELAN 1.2 for Linux on the recorded video. Location was annotated at the room level by two annotators. The audio part of the corpus was transcribed using Transcriber[6] by 5 annotators while the semantic annotation of the voice commands for Natural Language Understanding (NLU) was performed by 3 annotators using a home made web-based tool.

VI. EXPERIMENTS UNDERTAKEN WITH THE CORPORA

To illustrate the interests of the corpus, we present below experiments on two challenging tasks that need to be addressed for successful development of context aware voice-based smart home systems: Voice Activity Detection (VAD) and multi-resident localization. VAD aims at isolating speech signals from the ambient noise and is used as a front-end component in voice-based applications, such as automatic speech recognition, natural language understanding, or even paralinguistic information retrieval, e. g., emotion sensing or health monitoring. Resident localization aims at determining the physical (i. e., coordinate) or semantic location (e. g., room) of a resident in a home from sensor data. It is an essential component of many tasks including activity recognition and human-computer interaction (e. g., situated dialogue or prompting).

[6]http://trans.sourceforge.net/

A. Voice Activity Detection

A typical VAD task can be seen as a binary classification problem – speech vs. non-speech – performed time-continuously on the acoustic signal. Early approaches to VAD relied on simple energy thresholds or rules based on pitch and zero-crossing rate [23]. Whereas those methods perform well when there is little or no background noise, the performance degrades severely when spectral characteristics similar to speech are present in the background, even with the use of more sophisticated signal representations like auto-regressive (AR) model parameters [24]. For such challenging conditions of highly corrupted speech in real-life situations, data-driven methods such as supervised machine learning have shown promising results [25], especially when using long-span features [26], because the decision performed on each frame can be done in the context of the previous frames. A natural extension of this approach consists in the use of Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [27], because they can learn long-range contextual dependencies from the acoustic features by using dedicated memory cells.

VAD is performed on the VocADom@A4H corpus with the same LSTM-RNN model that was proposed in [27]. The open-source openSMILE toolkit [28] is used for running the features extraction from the acoustic signals which are then fed to the LSTM-RNN model to infer the level of voice activity. Standard RASTA-PLP with cepstral coefficients 1–18 and their first order derivate are extracted from the acoustic signal (frame size is 25 ms and hop size is 10 ms) and used as input of the networks after a z-normalization (zero mean, unit variance); the 0-th cepstral coefficient is voluntarily ignored to have the networks being invariant to the input level. The topology of the networks has an input layer matching the size of the RASTA-PLP features set, one recurrent hidden layer (four blocks with 50 LSTM cells each), and an output layer with a single linear unit. The VAD is thus performed as a regression task with the networks trained as regressors to output a voicing score for every frame in the range $[-1; +1]$; +1 indicating voicing, -1 indicating silence or noise.

The LSTM-RNN was trained and optimized on a large amount of labeled and diverse speech data that were synthetically generated by building random utterance sequences overlaid with additive noise. The Buckeye corpus [29], which consists of 26 h of spontaneous speech from 40 speakers collected in informal interview situations, and the TIMIT corpus [30], which includes 5.4 h of read speech data, were used as training, validating and testing materials. Four types of noise (babble, city, white and pink noise) were mixed with the speech data with various levels of SNRs [27]. Additionally, the full English audio tracks of four Hollywood movie DVDs were used as a second test set in challenging real-life conditions.

The area under receiver operating characteristic (ROC) curves (AUC) is used as evaluation metric. As in the original method [27], the thresholded VAD predictions are smoothed with a silence hysteresis of five frames, i. e., non-speech segments shorter than five frames are joined with adjacent speech segments. Results showed that, the VAD based on the LSTM-RNN model clearly outperformed all tested baseline algorithms [27], with the set of Hollywood movies being more challenging ($AUC = .722$) compared to the Buckeye and TIMIT corpora ($AUC = .961$).

TABLE IV
VAD RESULTS ON THE VOCADOM@A4H TEST SET FOR THE DIFFERENT PHASES OF THE EXPERIMENT AND THE PROCESSED SPEECH SOURCES. PERFORMANCE IS MEASURED USING AUC. BEST RESULTS OBTAINED FROM THE 16-CHANNEL ONLY ARE HIGHLIGHTED IN BOLD.

Source	Phase 1 (Elicitation)	Phase 2 (Visit)	Phase 3 (Lecture)	All
Headset	.782	.784	.805	.796
Arrays – averaged	.702	.651	.661	.659
Arrays – loudness	.690	.644	.660	.657
Arrays – LSTM-RNNs	**.704**	**.659**	**.675**	**.670**
Arrays – location	.703	.704	.664	.671

Whereas those corpora include a single acoustic signal, the VocADom@4H corpus features 16 audio channels that were recorded simultaneously in four different locations of the apartment, and which need to be processed dynamically in order to cope with the movements of the user. Different strategies were used to investigate the impact of this particular condition on VAD: 1) *array–location*, the array is selected for each frame according to the annotated location of the user, 2) *averaged*, all arrays are averaged into a single acoustic signal, 3) *loudness*, the loudest array is selected for each frame, and 4) *LSTM-RNNs*, the array with the highest voicing score inferred by the networks is selected for each frame. Table IV shows the AUC obtained for each phase of the experiment and compared with the headset microphone. Unsurprisingly, the baseline VAD with the headset microphone stays stable (between .782 – .805) since it is marginally perturbed by ambient noise. When the 16-channel data are considered, the distant speech condition involves a decrease in performance particularly strong for phase 2 (two-resident phase) and 3 (domestic noise). It is worth recalling that phase 2 was composed of noise and spontaneous speech while phase 3, although very noisy, contained only read speech. This explains the slightly better performance of phase 3. Among the 16-channel only informed methods, LSTM-RNNs exhibits clear highest performance. However, the location informed method (*array–location*) shows slightly better performance in phase 2 which is the multi-residents phase. This suggests that user location information from non-acoustic source can be successfully exploited to perform VAD. This kind of research can be further pursued using the VocADom@A4H corpus.

B. Multi-Sensor Multi-Resident Localization

Resident localization in home is usually performed either by worn sensors, video cameras or home automation sensors [31], [32]. Although worn sensors and video cameras can be used in certain circumstances, they can be too intrusive in daily life. Hence, sensor based approaches have been proposed [31], [33] but these have mostly emphasized the difficulty of the task. The accuracy decreases dramatically when the number

TABLE V
LOCALIZATION CLASSIFICATION RESULTS ON THE VOCADOM@A4H TEST SET; PERFORMANCE IS GIVEN IN % WEIGHTED AVERAGED F-MEASURE.

Sensor set	C4.5			seq2seq		
	Count	Dweller	Visitor	Count	Dweller	Visitor
PIR only	57.7	35.2	43.6	58.9	39.7	52.3
PIR+mic.	67.9	53.7	59.9	68.4	43.8	62.7

of dwellers become more than one (around 40-60% accuracy) even in case of dense set of sensors.

In this experiment we performed multi-resident localization (2 dwellers) and resident counting using only binary movement detectors (Passive Infra-Red sensors or PIR) and continuous microphone signals. A sequence-to-sequence model with attention [34] was used. In this model, the input $x = \{x_{t-(n-1)}, \ldots, x_t\}$ is composed of a sequence of feature vectors from time $t-(n-1)$ to t (current time) where n is the sequence length while the output $y = \{count, loc1, loc2\}$ is a sequence composed of the number of dwellers, the room location of dweller 1 and of dweller 2 at time t. The model uses a single RNN layer of gated recurrent unit (GRU) of size 128, both for the input and the output with an embedding layer of size 50. RNN based model had become very competitive in video-based tracking [35] but have still not be used in sensor-based localization. Sequence length was set to $n = 10$.

The VocADom@4H data set was partitioned into three speaker independent partitions: training set (S00-06, 5 men - 2 women), development set (S08, S10, 1 man - 1 woman), and test set (S07, S09, 1 man - 1 woman), with the latter being used for reporting the results. Feature vectors were computed using a sliding window. For the six binary sensors (e.g., infra-red motion detectors, switches), the number of firings in a time window was computed. For each of the 16 acoustic channels, the Root Mean Square (RMS) was computed in a time window and averaged per room resulting in 4 values per windows. Regarding location, the highest percentage of time of occupation of each room was computed for each time window and taken as the ground truth. As a result, 5962 instances were computed from the 11 records using a temporal window of size 15 seconds with 50% overlap. The longest location of the two main occupants of the smart home were computed for each window. Overall 56% of the time the smart home is occupied by one person, 39% by two people and 5% by nobody. For each of the two main occupants the following locations are considered : `kitchen`, `living room`, `outside`, `staircase`, `bedroom`, `entrance`, `bathroom`, `corridor`.

Table V presents the weighted F-measure score[7] for the task of resident counting and dweller and visitor localization. A decision tree (C4.5) was used as the instance based model of reference. Results show that Passive Infra-Red sensors alone (row `PIR`) do not provide enough information for a correct localization of the dweller. However, adding the acoustic information (`PIR+mic.`) does improve F-measure for both models. Performances of the DNN model are not has high as expected. However, it is well know that DNN learning requires a large number of data, a requirement which is not met in this study (only 3600 instances in the training set). Nevertheless, the seq2seq model performed the 3 tasks in only one model (while it necessitated three C4.5 models). Furthermore, the seq2seq model exhibited great consistency between the tasks of counting and localization since the inferred number of residents was always consistent with the inferred location (i.e., when only one resident is counted, there is always one resident predicted as being out).

[7]The weighted averaged F-measure takes class distribution into account.

Localizing one resident using home automation sensors and without worn sensors has been investigated for twenty years [36], [37], but it is only recently that the real-life problem of tracking *multiple* residents with only sensors firing has been considered [31], [33], [38]. The problem becomes much more difficult than in the single resident case since each sensor reading cannot be directly associated to a specific resident. Furthermore, contrary to video- or wifi-based localization, sensor readings contain far less redundant information to perform such tracking in real-time. The experiments exposed in this paper confirm the complementary role of standard home automation sensors and microphones and should orient future research towards probabilistic sequential models (e.g., CRF, HMM, RNN) and deep models which were difficult to start due to the shortage of data. Furthermore, the audio channel has been under-exploited in the pervasive community field while it can provide transient but accurate localization of a speaker [39]. This pleads for future research considering the whole information available at hand, which the VocADom@A4H corpus can support with its rich set of data.

VII. CONCLUSION AND FURTHER WORK

This paper summarizes the effort developed in the VOCADOM [40] project to collect a corpus of users' interaction in a Smart Home to be made available to the community. This is one of the few corpora that contain acoustic data *and* home automation sensor data. The protocol has been conceived to match two opposite constraints of control and ecology in the dataset, to involve several users and to let users behave as naturally as possible. This corpus is a valuable resource to perform studies related to human activity recognition, context aware interaction, decision-making, robust distant speech recognition applied to home automation control. Experiments performed on multichannel speech data and on home automation sensors for Voice Activity Detection and multi-resident localization as well as natural language understanding [41], [42] show the interest of the corpus to develop context-aware voice-based smart home control.

ACKNOWLEDGEMENTS

This work is part of the ANR VocADom project (ANR-16-CE33-0006). Part of experiments have been supported by the Amiqual4Home Equipex ("Investissements d'Avenir" program, ANR-11-EQPX-0002). The authors would like to thank the participants to who accepted to perform the experiments.

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] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, and P. Chahuara, "Evaluation of a context-aware voice interface for Ambient Assisted Living: qualitative user study vs. quantitative system evaluation," *ACM Transactions on Accessible Computing* , vol. 7, no. issue 2, pp. 5:1–5:36, May 2015.

[3] M. Weiser, "The computer for the 21st century," *Scientific American*, vol. 265, no. 3, pp. 66–75, 1991.

[4] G. M. Youngblood and D. J. Cook, "Data mining for hierarchical model creation," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 37, no. 4, pp. 561–572, 2007.

[5] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, "Accurate activity recognition in a home setting," in *Proceedings of UbiComp '08*, 2008.

[6] D. J. Cook and M. Schmitter-Edgecombe, "Assessing the quality of activities in a smart environment," *Methods of Information in Medicine*, vol. 48, no. 5, pp. 480–485, 2009.

[7] N. Zouba, F. Bremond, M. Thonnat, A. Anfosso, E. Pascual, P. Mallea, V. Mailland, and O. Guerin, "A computer system to monitor older adults at home: Preliminary results," *Gerontechnology Journal*, vol. 8, no. 3, pp. 129–139, July 2009.

[8] J. Cumin, G. Lefebvre, F. Ramparany, and J. L. Crowley, "A Dataset of Routine Daily Activities in an Instrumented Home," in *UCAmI 2017 - 11th International Conference on Ubiquitous Computing and Ambient Intelligence*, vol. 10586, Philadelphie, United States, 2017, pp. 413–425.

[9] T. L. M. van Kasteren, G. Englebienne, and B. J. A. Kröse, "Transferring knowledge of activity recognition across sensor networks," in *Proceedings of the 8th International Conference on Pervasive Computing*, 2010, pp. 283–300.

[10] H. Alemdar, H. Ertan, O. D. Incel, and C. Ersoy, "Aras human activity datasets in multiple homes with multiple residents," in *7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, 2013, pp. 232–235.

[11] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The CHiME challenges: Robust speech recognition in everyday environments," in *New Era for Robust Speech Recognition - Exploiting Deep Learning*. Springer, Nov. 2017, pp. 327–344.

[12] E. Principi, S. Squartini, F. Piazza, D. Fuselli, and M. Bonifazi, "A distributed system for recognizing home automation commands and distress calls in the italian language," in *Proceedings of Interspeech 2013*, Lyon, France, 2013, pp. 2049–2053.

[13] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, 2015, pp. 275–282.

[14] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, E. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury, and E. Jamet, "A French corpus for distant-microphone speech processing in real homes," in *Proceedings of Interspeech 2016*, 2016, pp. 2781–2785.

[15] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 32–36.

[16] S. Intille, K. Larson, E. Tapia, J. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson, "Using a Live-In Laboratory for Ubiquitous Computing Research," in *Pervasive Computing*, 2006.

[17] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A multimodal corpus recorded in a health smart home," in *Proceedings of LREC Workshop Multimodal Corpora and Evaluation*, Malta, 2010, pp. 99–105.

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

[19] P. Lago, F. Lang, C. Roncancio, C. Jiménez-Guarín, R. Mateescu, and N. Bonnefond, "The ContextAct@A4H real-life dataset of daily-living activities Activity recognition using model checking," in *CONTEXT*, ser. LNCS, vol. 10257, 2017, pp. 175–188.

[20] F. Aman, "Automatic speech recognition for ageing voices in the context of assisted living," Ph.D. dissertation, Université de Grenoble, 2014.

[21] S.-y. Takahashi, T. Morimoto, S. Maeda, and N. Tsuruta, "Dialogue experiment for elderly people in home health care system," in *Text, Speech and Dialogue*, Berlin, Heidelberg, 2003, pp. 418–423.

[22] S. Möller, F. Gödde, and M. Wolters, "Corpus analysis of spoken smart-home interactions with older users." in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.

[23] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *IET Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[24] S. Mousazadeh and I. Cohen, "AR-GARCH in Presence of Noise: Parameter Estimation and its Application to Voice Activity Detection," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 4, pp. 916–926, 2011.

[25] A. Misra, "Speech/nonspeech segmentation in web videos," in *Proceedings of Interspeech 2012*. ISCA, 2012.

[26] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselý, and P. Matějka, "Developing a speech activity detection system for the DARPA RATS program," in *Proceedings of Interspeech*, 2012.

[27] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies," in *Proc. of INTERSPEECH*, 2013.

[28] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia (ACM MM)*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[29] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA, 2007.

[30] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgrena, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993.

[31] A. S. Crandall and D. J. Cook, "Tracking systems for multiple smart home residents," in *Human behavior recognition technologies: Intelligent applications for monitoring and security*, 2013, pp. 111–129.

[32] S. A. Mehdi and K. Berns, "A survey of human location estimation in a home environment," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 135–140.

[33] T. Miyazaki and Y. Kasama, "Multiple human tracking using binary infrared sensors," *Sensors*, vol. 15, no. 6, pp. 13 459–13 476, 2015.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[35] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[36] C. R. Wren and E. M. Tapia, "Toward scalable activity recognition for sensor networks," in *Location- and context-awareness*, 2006.

[37] P. Chahuara, F. Portet, and M. Vacher, "Location of an Inhabitant for Domotic Assistance Through Fusion of Audio and Non-Visual Data," in *Pervasive Health*, 2011, pp. 1–4.

[38] A. Benmansour, A. Bouchachia, and M. Feham, "Multioccupant activity recognition in pervasive smart home environments," *ACM Comput. Surv.*, vol. 48, no. 3, pp. 1–36, 2015.

[39] S. Sivasankaran, E. Vincent, and D. Fohr, "Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment," in *Proceedings of Interspeech 2018*, Hyderabad, India, 2018.

[40] M. Vacher, E. Vincent, M.-E. Bobillier Chaumon, T. Joubert, F. Portet, D. Fohr, S. Caffiau, and T. Desot, "The VocADom Project: Speech Interaction for Well-being and Reliance Improvement," in *MobileHCI 2018 - workshop Designing Speech and Language Interactions for Mobiles and Wearables*, Barcelona, Spain, 2018.

[41] T. Desot, S. Raimondo, A. Mishakova, F. Portet, and M. Vacher, "Towards a French Smart-Home Voice Command Corpus: Design and NLU Experiments," in *Proceedings of 21st International Conference on Text, Speech and Dialogue TSD 2018*, 2018, pp. 509–517.

[42] A. Mishakova, F. Portet, T. Desot, and M. Vacher, "Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in Smart Homes," in *PerDial 2019*, Kyoto, Japan, 2019.