Speech and Speaker Recognition for Home Automation: Preliminary Results

Michel Vacher, Benjamin Lecouteux, Javier Serrano-Romero, Moez Ajili, François Portet, Solange Rossato

Speech and Speaker Recognition for Home Automation: Preliminary Results

Michel Vacher, Benjamin Lecouteux, Javier Serrano Romero,
Moez Ajili, François Portet and Solange Rossato
CNRS, LIG, F-38000 Grenoble, France
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
Laboratoire d'Informatique de Grenoble, GETALP Team
Grenoble, France
{Michel.Vacher, Benjamin.Lecouteux}@imag.fr
{ Javier.Serrano-Romero, Moez.Ajili }@imag.fr
{ Francois.Portet, Solange.Rossato}@imag.fr

***Abstract*—In voice controlled multi-room smart homes ASR and speaker identification systems face distance speech conditions which have a significant impact on performance. Regarding voice command recognition, this paper presents an approach which selects dynamically the best channel and adapts models to the environmental conditions. The method has been tested on data recorded with 11 elderly and visually impaired participants in a real smart home. The voice command recognition error rate was 3.2% in off-line condition and of 13.2% in online condition. For speaker identification, the performances were below very speaker dependant. However, we show a high correlation between performance and training size. The main difficulty was the too short utterance duration in comparison to state of the art studies. Moreover, speaker identification performance depends on the size of the adapting corpus and then users must record enough data before using the system.**

***Index Terms*—Home Automation, Voice controlled smart home, Vocal command, Speaker recognition.**

I. Introduction

With the ageing of the population, the issues and challenges created by homecare and the increase loss of autonomy is a major concern over the coming years in European Union (EU). Anticipating and responding to the needs of persons with loss of autonomy with Information and communications technology (ICT) is known as ambient assisted living (AAL). In this domain, the development of Smart Homes is seen as a promising way of achieving in-home daily assistance [1], [2], [3] because one of the first wishes of this population is to live in their own home as comfortably and safely even if their autonomy decreases. Efforts are needed for the building industry to put into practice smart home technologies that support independent living [4] because some sensors and Ambient Intelligence devices must be set up in the home.

Some surveys listed the sensors which are involved in this framework [5], [6] and the most recent of them include microphones and speech analysis [7] but they brought to light the challenges that must be taken up before putting into practice audio analysis in real home [8]. These challenges include: 1) distant speech and noisy conditions [9], [10], 2) hands-free interaction, 3) affordability by people with limited financial means, 4) real-time interaction, 5) respect of privacy, but it should be noted that the intrusiveness of an assistive technology can be accepted if the benefit is worth it. Indeed, speech technologies are well adapted to people with reduced mobility or some disabilities, and who are not familiar with technical devices and as well as some emergency situations because distant interaction is possible.

Nowadays, voice–user interfaces (VUIs) are frequently employed in close application domains (e.g., smart phone) and are able to provide interaction using natural language so that the user does not have to learn complex computing procedures [11]. Some industrial products are available, such ase Amazon Echo [12], but they use a cloud-based processing system that could become a problem related to privacy.

A large number of projects were related to assistive technologies such as CASAS [13], AGING IN PLACE [14], DESDHIS [15], GER'HOME [16] and SOPRANO [17]. Moreover, there is a rising number of a smart home projects or studies that considers speech recognition in their design, COMPANIONABLE [18], PERS [19], Filho et al. [20], SWEET-HOME [21] and CIRDO [22]. Some of them is being focused on pathologic voices (i.e., Alzheimer) like ALADIN [23], HOMESERVICE [24], and PIPIN [25]. The aim of DIRHA [26] is the study of distant speech recognition at home. However, such technologies must be validated in real situations and SWEET-HOME [27]SWEET-HOME [27] is, to the best of our knowledge, the first vocal command system which was evaluated online in a real smart home with potential users [28].

In this paper, we present an approach to provide voice commands in a multi-room smart home for seniors and people with visual impairments. Our aim is both to recognize, thanks to speech analysis, the home automation command and to identify the speaker. Indeed, before undertaking an action, it is necessary to have a good knowledge regarding the context [29]: the location of the person in the housing, her activity, her identity for permission and preferences. In our approach, we

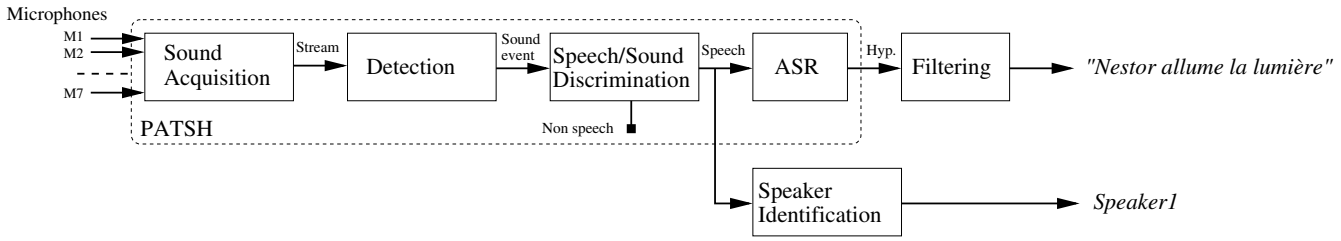

Fig. 1. General architecture of the audio processing system.

address the problem by using several mono-microphones set in the ceiling, selecting the "best" source and employing a pattern matching approach at the output of the speech recognition system. This approach has been chosen against a noise source separation approach which can be computationally expensive, (i) is sensitive to sample synchronization problem (which cannot be assumed with cheap non professional devices) and (ii) is still not solved in real uncontrolled conditions. Hands-free interaction is ensured by constant keyword detection.

Indeed, the user must be able to control the environment without having to wear a specific device for physical interaction (e.g., a remote control too far from the user when needed). Though microphones in a home are a real breach of privacy, by contrast to current smart-phones, we address the problem by using an in-home ASR engine rather than a cloud based one (private conversations do not go outside the home). Moreover, the limited vocabulary ensures that only speech relevant for the command of the home is correctly decoded. Finally, a strength of the approach is to have been evaluated with real users in realistic conditions.

The paper is organized as follow. Section II presents the method implemented for spoken command and speaker recognition in the home. Section III presents the experimentation and the results which are discussed in Section IV.

II. Method

The multi-source voice command recognition is to be performed in the context of a smart home which is equipped with microphones, one or two microphones are set into the ceiling of each room, as shown Figure 2. The audio processing task is both to recognize predefined sentences that correspond either to a home automation command or to a distress call, and to identify the speaker. The audio module should not process other private conversations. Once a command is recognized (e.g., "turn on the light"), it is sent to a intelligent controller which manages the home automation system (e.g., light up the lamp the closest to the person). For more information about the home automation management, the reader is refereed to [29].

The general architecture of the audio processing system is shown in Figure 1. For real-time audio analysis the PATSH framework was developed to manage the acquisition/processing flow of the sound events detected. In PATSH, several processors are plugged in and synchronised, such as the sound and speech analysis modules and the multichannel

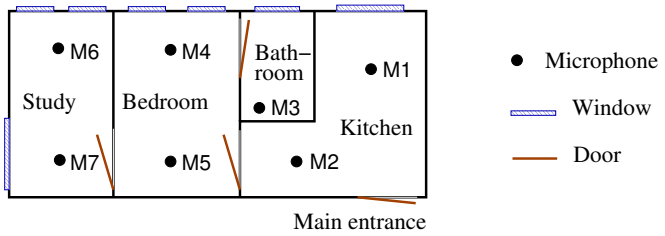

Fig. 2. Position of the microphones in the smart home.

data acquisition card. The microphone data is continuously acquired and sound events are detected on the fly by using a wavelet decomposition and an adaptive thresholding strategy [30]. Sound events are then classified as non-speech or speech.

Non-speech sounds are ignored while speech sounds are sent to an ASR system [31] and to the Speaker Identification system described in Section II-B. The answers of the ASR and of the Speaker Identification system are then sent to the intelligent controller, which, thanks to the context (the other information available through the home automation system), makes a context aware decision. In the remaining of this section, we describe the ASR system in Section II-A and the speaker recognition system in Section II-B.

We focus on the ASR system and present different strategies to improve the recognition rate of the voice commands in Section II-A. In the second instance, we focus on speaker identification as described in Section II-B.

A. Vocal command recognition

For the reason emphasized in the introduction, the methods do not concentrate on enhancement of the signal but on the decoding of data output from the channel with the best Signal to Noise Ratio ("*best channel*") and on the use of *a priori* information at the language level to generate the hypothesis which is consistent with the task. Therefore, the methods employed at the acoustic and decoding level are presented in this part in Sections II-A1, II-A2 and II-A5. The Spoken Keyword Spotting method is presented as baseline in Section II-A4.

1) The acoustic modeling: The Kaldi speech recognition tool-kit [32] was chosen as ASR system. Kaldi is an open-source state-of-the-art ASR system with a high number of tools and a strong support from the community. In the experiments, the models are context-dependent three-state left-right HMMs. Acoustic features are based on mel-frequency cepstral

coefficients, 13 MFCC-features coefficients are first extracted and then expanded with delta and double delta features and energy (40 features). Acoustic models are composed of 11,000 context-dependent states and 150,000 Gaussians. The state tying is performed using a decision tree built from phonetical information. In addition, acoustic adaptation is performed thanks to off-line Feature space Maximum Likelihood Linear Regression (fMLLR) linear transformation.

The acoustic models were trained on 500 hours of transcribed French speech composed of the ESTER 1&2 (broadcast news and conversational speech recorder on the radio) and REPERE (TV news and talk-shows) challenges as well as from 7 hours of transcribed French speech from 60 speakers interacting in the Smart home [33], called SH (SWEET-HOME) in the text. This corpus is made of the Multimodal subset (27 speakers, age: 22-63, 7 female&14 male), the Home Automation subset (23 speakers, age: 19-64, 9 female&14 male) and the Interaction subset (16 speakers, age: 19-62, 7 female&9 male).

2) Subspace GMM Acoustic Modelling: The GMM and Subspace GMM (SGMM) both model emission probability of each HMM state with a Gaussian mixture model, but in the SGMM approach, the Gaussian means and the mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections. The SGMM model [32] is described in the following equations:

$$\begin{cases} p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i), \\ \mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \\ w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}. \end{cases}$$

where $\mathbf{x}$ denotes the feature vector, $j \in \{1...J\}$ is the HMM state, i is the Gaussian index, m is the substate and c_{jm} is the substate weight. Each state j is associated to a vector $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S is the phonetic subspace dimension) which derives the means, μ_{jmi} and mixture weights, w_{jmi} and it has a shared number of Gaussians, I. The phonetic subspace $\mathbf{M}_i$, weight projections $\mathbf{w}_i^T$ and covariance matrices $\mathbf{\Sigma}_i$ i.e; the globally shared parameters $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$ are common across all states. These parameters can be shared and estimated over multiple recording conditions.

A generic mixture of I Gaussians, denoted as Universal Background Model (UBM), models all the speech training data for the initialization of the SGMM.

Our experiments involved obtaining SGMM shared parameters using both SWEET-HOME data (SH 7h) and clean data (500h). In the GMM system, the two training data set are just merged in a single one. We propose to train 2 UBMs:

- The first one is a classical SGMM system using all the data to train the UBM (1K gaussians). In the experiments, this SGMM model is named **SGMM1**.
- For the second one, two UBM are trained respectively on SWEET-HOME data and clean data. The two obtained UBMs contain 1K gaussians and are merged into a single one mixed down to 1K gaussian (closest Gaussians pairs are merged [34]) , this SGMM is named **SGMM2**. The aim is to specifically bias the acoustic model with the smart home conditions.

3) The language models: For the decoding, a 2-gram language model (LM) with a 10K lexicon was used. It results from the interpolation of a generic LM and a specialized LM. The *generic* LM was estimated on about 1000M of words from the French newspapers *Le Monde* and *Gigaword*, and the broadcast news manual transcripts provided during the ESTER campaign. The *specialized* LM was estimated from the grammar and from the transcript of the 60 speakers, containing voice commands and casual speech.

4) Spoken Keyword Spotting: The problem of recognizing voice commands with a predefined grammar but not other conversation can be seen as a spoken keyword spotting problem. Given the uncertainty of the ASR system output, spoken keyword spotting has mainly been addressed by searching instances of particular keywords in the ASR set of hypotheses or lattice obtained after processing an utterance [35]. In this work, we use the method of Can and Saraçlar [36] for Spoken Term Detection (STD) from the ASR decoding lattice of an utterance. In this approach, the lattice is transformed into a deterministic weighted Finite State Transducer (FST) called Timed Factor Transducer (TFT) embedding informations for the detection (utterance ID, start and end time and posterior score). A search of a string X in a speech utterance is then a composition of the automaton representation of X and the TFT to give the resulting transducer R which contains all the possible successful detections and their posterior probability. The interest of such approach is the fact that search complexity is linear and that performing several searches in a same utterance can be done with the same TFT. Moreover, the FST formalism makes filtering with a predefined grammar easy by using the composition operator.

5) Detection of voice commands: We propose to transcribe each voice command and ASR output into a phoneme graph in which each path corresponds to a variant of pronunciation. For each phonetized ASR output T, every voice commands H is aligned to T using Levenshtein distance. The deletion, insertion and substitution costs were computed empirically while the cumulative distance $\gamma(i,j)$ between H_j and T_i is given by Equation 1.

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

The voice command with the aligned symbols score is then selected for decision according a detection threshold. This approach takes into account some recognition errors like word endings or light variations. Moreover, in a lot of cases, a miss-decoded word is phonetically close to the true one (due to the close pronunciation). From this the DER (Domotic Error Rate i.e., home automation error rate) is defined as:

$$\text{DER} = \frac{\text{Missed} + \text{False Alarms}}{\text{Voice Commands}_{\text{syntactically correct}}} \quad (2)$$

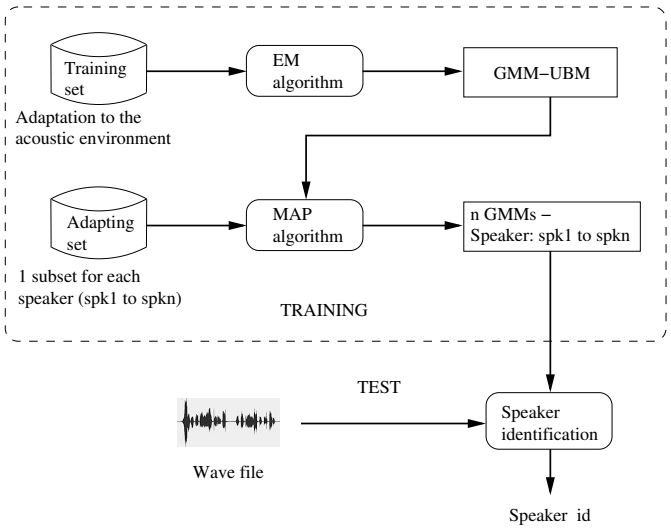

Fig. 3. General architecture of the speaker identification system.

For the DER, the ground truth is the number of uttered voice commands respecting the grammar. I.e., the utterances where the person's intention was to utter a vocal command but was not following the voice command syntax were not considered as a true voice commands. The Missed correspond to the true voice commands not recognized and the False Alarms to sound events incorrectly classified as voice commands.

B. Speaker recognition

As emphasized in the introduction, speaker identification gives complementary information for context inference (location, activity, identity of speaker who is uttering a command). Figure 3 presents the speaker identification method; it is divided into two phases:

- Training phase in which a large data set is used to build an UBM that models the whole acoustic space. Then, speaker models are obtained by UBM adaptation.
- Testing phase in which a speech occurrence will be assigned to one of the existing speakers. The used UBM is modeled by GMMs.

1) UBM: UBM is trained by Expectation-Maximisation (EM) algorithm [37] which is divided in two steps:

- Expectation evaluation (E), where the likelihood expectation is calculated using the last observed values.
- Maximisation step (M), where the parameters maximum likelihood are estimated by maximizing the likelihood calculated in the E step.

The EM algorithm is iterative. Parameters estimated in the Maximisation (M) step are used as a check point to start a new expectation evaluation, by applying an iteration. Between two successive iterations, the algorithm guarantees the growth of a likelihood objective function parameters given x. We should notice that EM converges to a local maximum.

2) Adaptation to the speaker: We derive the speaker model by adapting the parameters of the UBM using the speaker's training speech and a form of Bayesian adaptation [38]. Unlike the standard approach of maximum likelihood training of a model for the speaker independently of the UBM, the basic idea in the adaptation approach is to derive the speaker's model by updating the well-trained parameters in the UBM via Maximum A Posteriori adaptation (MAP). Like the EM algorithm, the adaption is a two step estimation process. The first step is identical to the expectation step of the EM algorithm, where estimates of the sufficient statistics of the speaker's training data are computed for each mixture in the UBM. Unlike the second step of the EM algorithm, for adaptation these new sufficient statistic estimates are then combined with the old sufficient statistics from the UBM mixture parameters using a data-dependent mixing coefficient. Typically, we only adapt the means of the Gaussian, keeping the variance unchanged.

3) Scoring: The system proposed by Jousse et al. [39] is based on Semantic Classification Trees (SCT). For each utterance, a score is calculated indicating the degree of support to every speaker. Then, the speaker who has the maximum score is identified. A detailed description of this system is presented in [40] and [41].

III. Experimentation and results

A. Vocal Command Recognition: Live Experiment

An experiment involving elderly and visually impaired people was run in the DOMUS smart home which is part of the experimentation platform of the LIG laboratory. This experiment is fully described in [33] and only the aspects essential for the comprehension of the reader are recalled in this article. This smart home is a four-room flat (see Figure 1) equipped with home automation system and with 7 microphones set in the ceiling for audio analysis. A communication device was also present to allow video-conferencing. The SWEET-HOME system consisted in the PATSH software presented in Section II which was continuously analysing the audio streams to detect voice commands [42] and an intelligent controller [29] in charge of executing the correct action (e.g., lighting the lamp, or giving the temperature using TTS) based on these voice commands.

Each participant had to follow 4 successive scenarios with the following topics: 1) 'finishing breakfast and going out', 2) 'coming back from shopping and cleaning', 3) 'communicating with a relative', and 4) 'waiting for friends'. Each of these scenarios was designed to last between 5 to 10 minutes but there was no constraint on the execution time. Scenario 1 and 2 were designed to have the user performing daily activities while uttering voice commands. The participant was provided with a list of actions to perform and voice commands to utter. Each participant had to use vocal commands to turn the light on or off, open or close blinds, ask about temperature and ask to call his or her relative.

Six seniors and five people with visual impairments were recruited. The seniors (81.2 years old (Standard Deviation: SD=5.8)) were women living alone in an independent non-hospitalised accommodation. The focus of study was to target

TABLE I
SPEECH AUDIO DATA RECORDED DURING THE SCENARIOS (FOR EACH SPEAKER, A READ TEXT WAS RECORDED FOR SPEAKER ADAPTION)

ID	Category	Age	Sex	Scenario duration	Speech utterances	Voice commands		SNR (dB)
						uttered	missed	
S01	Aged	91	F	24mn	59	37	30	16
S02	Visually	66	F	17mn 49s	67	26	5	14
S03	Visually	49	M	21mn 55s	53	26	9	20
S04	Aged	82	F	29mn 46s	74	27	12	13
S05	Visually	66	M	30mn 37s	47	25	10	19
S06	Aged	83	F	22mn 41s	65	31	20	25
S07	Aged	74	F	35mn 39s	55	25	13	14
S08	Visually	64	F	18mn 20s	35	22	8	21
S09	Aged	77	F	23mn 5s	46	23	17	17
S10	Visually	64	M	24mn 48s	49	20	7	18
S11	Aged	80	F	30mn 19s	79	26	18	23
All	-	-	-	4h 39mn	629	291	149	-

seniors who where on the edge of losing some autonomy and not seniors who have lost complete autonomy. In other words, we sought seniors who were still able to make a choice regarding how the technology could help them in case of any physical degradation. The visual impaired category (62.2 (SD=6.9) years old, 3 were women) was composed of adult people living either single or as couple and whose handicap was acquired after their childhood. No upper age limit was given. The participants were not completely blind but their visual acuity was very low.

1) Spoken commands: Possible vocal commands were defined using a very simple grammar which was built after a study revealing that targeted users prefer precise short sentences over more natural long sentences [11], this was corroborated in [28]. As shown below, each vocal command is either an emergency call or a command intended to control a device. Every command starts with an optional key-word (e.g. *'Nestor'*) to make clear whether the person is talking to the smart home or not. Some basic commands are then *'turn on the light'* or *'what is the temperature'*:

set an actuator on/off:	`key initiateCommand object` (e.g., `Nestor ferme fenêtre`) (*e.g., Nestor close the window*)
emergency call:	`key emergencyCommand` (e.g., `Nestor au secours`) (*e.g., Nestor help*)

2) Acquired data: At the beginning of the experiment, the participants were asked to read a short text (288 words in French), this text was used to adapt the acoustical models to the speaker before performing the scenarios.

During the experiment, audio data was recorded in two ways. Firstly, the 7-channel raw audio stream was stored for each participant for further analysis. Secondly, audio events were automatically extracted by the PATSH software on the fly. Some of the events were missed or discarded and some of the detected speech events were misclassified as everyday life sound, and some noise were misclassified as speech (bell ring, music, motor). In the later case, these non-speech events were sent to the ASR. A manual segmentation of the corpus was done thanks to the full records of the experiments using the Transcriber software [43]. Finally, these two speech data sets (manual vs. PATSH segmentation) were transcribed using Transcriber . For the PATSH data set, there are 617 uttered sentences. 291 were home automation commands (46%), 66 (10%) were actually generated by the speech synthesizer, 10 (2%) were noise occurrences wrongly classified as speech and 250 were other spontaneous speech (42%, mostly during the video-conferencing with a relative). Only 29 speech utterances were missed (4%), but 85 of the detected ones were rejected (14%) either because their SNR was below 0dB or because they were out of the acceptable duration range (2.2 seconds). Therefore, 18% of the utterances were not treated by the system. The recorded audio data are summarized in Table I.

B. Spoken command recognition: Off line experiments

The methods presented in Section II were run on the data set presented in Table I. Regarding S11, PATSH crashed in the middle of the experiment and due to time constraints S11 data was not considered in the study. Therefore 550 sentences (2559 words) including 250 commands, questions and distress calls (937 words) were used. Two acoustic models were used: AM (500h) and AM (500h+ SH, were SH = SWEET-HOME data), speaker adaptation was provided by fMLLR

using the text read by each speaker. The two versions of the Subspace Gaussian Mixture Models (SGMM1 and 2 cf. Section II-A2) were also applied to all different combinations. All the methods were run on the transcribed data (manually annotated) and on the PATSH data (automatically segmented). We present the keyword spotting approach in order to show that a conventional approach is limited because of language variations introduced by the protagonists (i.e. home automation commands are rarely pronounced correctly).

Results on manually annotated data are given Table III. The most important performance measures are the Word Error Rate (WER) of the overall decoded speech and those of the specific voice commands as well as the Domotic Error Rate (DER: c.f. Equation 2).

It can be noticed that most of the improvement is brought by the use of fMLLR and the use of data adapted to the acoustic environment (the SH dataset). The WER obtained from the overall speech goes from 59.8 to 35.7. But most of this reduction is driven by the dramatic decrease of error in the voice command decoding. It starts from 32.9% for the baseline, and is reduced to 22.8% with the use of an acoustic model adapted to the smart home and to 14.0% thanks to speaker adaptation. Best results, WER=10.1%, DER=3.2%, are obtained by using SGMM applied to the utterance with the highest SNR (best channel).

Regarding the data set extracted by PATSH, the original ASR performance with a decoding online in real time during the experiment in the smart home and on only one channel [42] was WER=43.2%, DER=41%. By using these same data, the automatically segmented corpus, and AM (500h+SH), SAT fMLLR, SGMM2 on the best channel or with a specific analysis system operating on 2 channels, we obtained the results presented in Table III. Results are somewhat lower than those obtained with the manually segmented data. The most important contribution to the DER is due to missed speech utterances at the detection or speech/sound discrimination level. Therefore this is a very significant improvement from the experimental condition.

TABLE II
WER AND DER FOR THE MANUALLY SEGMENTED DATA ON THE BEST CHANNEL

Method	WER all (%)	WER voice commands (%)	DER (%)
Keyword Spotting **SGMM2:** AM (500h + SH 7h), SAT + fMLLR	-	-	57.6
GMM: AM (500h)	59.8	32.9	20.8
GMM: AM (500h), SAT + fMLLR	46.0	15.9	5.6
GMM: AM (500h + SH 7h)	51.9	22.8	14.4
GMM: AM (500h + SH 7h), SAT + fMLLR	39.0	14.0	4.4
SGMM1: AM (500h+SH) SAT + fMLLR	38.1	11.4	3.6
SGMM2 : AM (500h+SH), SAT + fMLLR	36.1	10.9	3.2

C. Speaker Recognition

Acoustic features were of MFCC parameters, their derivatives, and 11 second order derivatives (the frequency window was restricted to 300-3400 Hz). A normalization file based process was applied, so that the distribution of each cepstral coefficient was 0-mean and 1-variance for a given utterance.

For modeling the GMM-UBM, ALIZE toolkit [44] was used. UBM is trained using 7,803 sessions from 24 speakers (using the speech corpus "Home Automation subset" of the Sweet-Home corpus [33] recorded by 24 speakers) trained by EM/ML algorithm (with a variance flooring ≈ 0) on about 2 millions of speech frames. It resulted in 64 Gaussian components.

For evaluation of the system, the "Interaction subset" of the Sweet-Home corpus [33] was used. This corpus is made of two separated parts and was recorded by the 11 specific speakers (5 visually impaired and 6 elderly people). Before participating to the experiment, each speaker read 21 vocal commands (i.e., 99 words) and a newspaper article (205 words) in the Domus smart home. These data are dedicated to speaker adaptation and they compose the training part of the corpus. Next, the same persons play the scenarios interacting with the Vocal Home Automation system in order to record the second part of the corpus. This part of the corpus is the test data set.

For MAP adaptation to each user, we used the training set and MAP algorithm that is implemented in the LIUM toolkit [39]. The duration of speech used to adapt speakers models was between 70 and 268s.

The test dataset is detailed detailed in Table IV. The speaker must be identified from a single sentence which represent a very short time, the average duration is 2.55s. Therefore, the speaker have to be identified among 11 possible speakers.

The performances are evaluated for each speaker in terms of Speaker Recognition Rate (SRR):

$$\text{SRR} = \frac{\text{Number of well identified utterances}}{\text{Total number of utterances}} \quad (3)$$

The recognition rate is given for each speaker Table V. The average recognition rate is 70% for all speaker, 75.6% for the visually impaired group and 62.6% for the elderly group.

TABLE III
WER AND DER FOR THE AUTOMATICALLY SEGMENTED DATA ON THE BEST CHANNEL

Method	Number of channels	WER all (%)	DER (%)
SGMM2: AM (500h + SH 7h), SAT + fMLLR	1	49.0	13.6
SGMM2: AM (500h + SH 7h), SAT + fMLLR	2	49.0	13.2

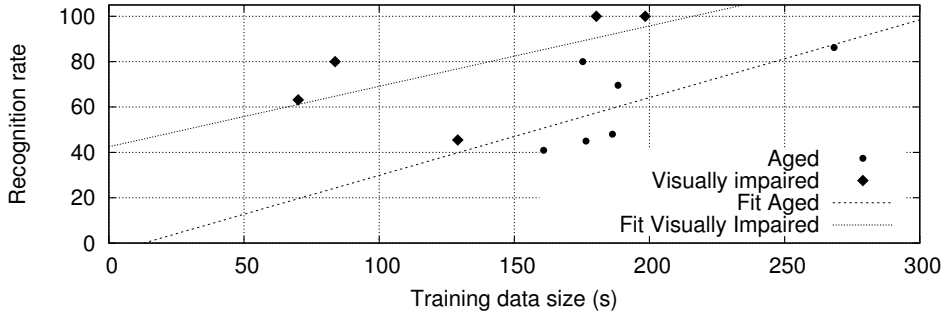

Fig. 4. Speaker recognition rate as a function of the size of the adaptation corpus for the two speaker groups.

Speaker	Average duration	Standard deviation
1	3.05 s	0.97 s
2	2.28 s	0.55 s
3	2.18 s	0.47 s
4	2.00 s	0.69 s
5	2.33 s	0.70 s
6	2.95 s	1.09 s
7	2.66 s	0.72 s
8	2.66 s	0.20 s
9	2.29 s	0.64 s
10	2.46 s	0.40 s
11	2.79 s	0.44 s
All	2.55 s	0.53 s

TABLE IV
AVERAGE DURATION OF A VOCAL COMMAND FOR THE HOME INTERACTION SUBSET.

Speaker	Genre	Type	SRR
1	F	Elderly	86.21 %
2	F	Visually impaired	45.45 %
3	M	Visually impaired	80.00 %
4	F	Elderly	40.91 %
5	M	Visually impaired	100.00 %
6	F	Elderly	48.00 %
7	F	Elderly	80.00 %
8	F	Visually impaired	63.16 %
9	F	Elderly	45.00 %
10	M	Visually impaired	100.00 %
11	F	Elderly	69.56 %

TABLE V
SPEAKER RECOGNITION RATE.

Input \ Identified	*Men*	*Women*
Men	48	1
Women	2	178

TABLE VI
CONFUSION MATRIX FOR GENDER RECOGNITION.

Figure 4 gives a representation of the performance for elderly and visually impaired speakers as a function of the size of the adapting corpus for the corresponding speaker. A curve fit is plotted for these two kinds of people, the 2 curves show a great difference between these 2 groups. In most cases and for elderly speakers, the performance is lower with a same size of training data as for visually impaired people. Therefore for this kind of population and before the installation in the home of the individual, it will be necessary in a first phase to record speech in unconstrained conditions.

The confusion matrix in table VI shows a good discrimination in gender, only 3 sentences out of 229 are not assigned to the right gender.

IV. DISCUSSION

Efficient on-line recognition of voice commands is mandatory for the dissemination of in-home VUI. This task must address many challenges such as distant speech recognition and respect for privacy. Moreover, such technology must be evaluated in realistic conditions. In this paper, we showed that a careful selection of the best channel as well as good adaptation to the environmental and acoustic characteristics increase dramatically the voice command classification performance. In the manual segmentation, SGMM acoustic model learned from data previously acquired in the smart home

as well as fMLLR diminish the DER from 20.8% to 3.2% surpassing more standard methods such as keyword spotting. In the recognition task based on the PATSH detection and discrimination, the best technique (2-channel SGMM, fMLLR) shows a rise of DER to 13.2%. This can be explained by the imperfect detection, segmentation and classification of the system. Indeed, some sentences were missing or split in two parts (e.g., 'Nestor light the lamp' $\rightarrow$ 'Nestor' then 'light the lamp'). Hesitations in real speech are natural but it is still unclear whether they are going to be frequent in real use or due to the experimental condition (people must learn the grammar).

Regarding the speaker identification, the presented results are not as good as those of the state of arts. Indeed, State of the Arts systems are configured for larger training data set and are evaluated on speech segments whose duration is about 30s while a home automation command lasts 2.55 seconds. For example, the STC Speaker Recognition System for the NIST i- Vector Challenge presented by Novosellov et al. [44] used segment durations with a mean of 39.58 seconds. Our system was configured in the same way. Nevertheless, our system shows good performances for male, evaluated to 93.3%. Our further research will be done to improve system precision by: (i) optimizing the training conditions, and (ii) using different process for male and female.

It can be noticed that Speakers 2, 4, 6 and 9 present the lowest SRR. Two explanations to this result could be proposed. Firstly, it is reasonable to think that speaker verification is very dependant to the enrollment and/or test conditions (utterance length, quality, channel mismatch, etc.). Secondly, this result could be linked directly to the "speaker factor". This issue has been discussed deeply in [45] and [46][46].

V. Conclusion

Studies presented in this paper are a part of a more general research project aiming at developing a system for home automation vocal commands which could be implemented in the personal home of elderly or frail people living alone. There are still important challenges to overcome before implementing this solution, particularly speech analysis in distant speech conditions, automatic speech recognition of elderly people or of expressive speech and more generally context aware interaction. Moreover, this new technology must be validated in real conditions with potential users, therefore our evaluations were conducted using a corpus recorded in a smart-home by potential users [28]. This paper presents and evaluates a system for vocal command recognition and a system for speaker identification aiming at delivering this information to an intelligent controller in charge of home automation driving.

Regarding spoken command recognition, performances are sufficient but there are still improvements to be performed before any use in real conditions. Indeed, in the experiment, people regularly deviated from the grammar (e.g., adding politeness terms or reformulation) and did not like the predefined chosen keyword. An interesting research direction would be to adapt the language model to the words a user 'naturally' utters in different situations, hence learning the grammar from the data rather than imposing a grammar. Another direction would be to exploit the smart-home capacity of sensing the environment to provide context-aware ASR. However, speaker recognition performances do not allow to identify the speaker who uttered a vocal command, especially if he is an elderly person, with a precision above 40%. This identification is an important information which can allow the home automation system to undertake the appropriate action and a future work will be related to determine the way of improvement of these first results. One possibility is to use information available from other sensors. For example, if it is established that there is only one person in the house and if a speaker was beforehand recognized, in that case the system could know that there is an error if another speaker is recognized. Thus the last data could be useful to serve as additional adaptation data.

Acknowledgement

This study was funded by the National Agency for Research under the project Sweet-Home (ANR-2009-VERS-011).

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] L. C. De Silva, C. Morikowa, and I. M. Petra, "State of the art of smart homes," *Engineering Applications of artificial Intelligence*, vol. 25, pp. 1313–1321, 2012.

[3] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen, and L. P. De Witte, "Literature review on monitoring technologies and their outcomes in independently living elderly people," *Disability and Rehabilitation: Assistive Technology*, pp. 1–24, 2014.

[4] N. Labonnote and K. Holand, "Smart home technologies that support independent living: challenges and opportunities for the building industry – a systematic mapping study," *Intelligent Buildings International*, 2015, 26 pages.

[5] Q. Lê, H. B. Nguyen, and T. Barnett, "Smart homes for older people: Positive aging in a digital world," *Future Internet*, vol. 4, pp. 607–617, 2012.

[6] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 579–590, 2013.

[7] Q. Ni, A. B. Garcia Hernando, and I. P. de la Cruz, "The elderly's independant living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development," *Sensors*, vol. 15, pp. 11 312–11 362, 2015.

[8] 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.

[9] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Published by Wiley, 2009.

[10] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013.

[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] "Amazon Echo," http://www.amazon.com/Amazon-SK705DI-Echo/dp/B00X4WHP5E, accessed: 2015-09-17.

[13] C. Chen and D. J. Cook, "Behavior-based home energy prediction." in *IEEE Intelligent Environments*, 2012, pp. 57–63.

[14] M. Skubic, G. Alexander, M. Popescu, M. Rantz, and J. Keller, "A smart home application to eldercare: Current status and lessons learned," *Technology and Health Care*, vol. 17, no. 3, pp. 183–201, 2009.

[15] A. Fleury, M. Vacher, and N. Noury, "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*, vol. 14, no. 2, pp. 274 –283, march 2010.

[16] 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.

[17] P. Wolf, A. Schmidt, and M. Klein, "Soprano - an extensible, open aal platform for elderly people based on semantical contracts," in *3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'08), 18th European Conference on Artificial Intelligence (ECAI 08), Patras, Greece*, 2008.

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

[19] 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, 2009.

[20] 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.

[21] 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, aug 2011, p. 4p.

[22] S. Bouakaz, M. Vacher, M.-E. Bobillier-Chaumon, F. Aman, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, E. Desserée, P. Traineau, J.-P. Vimon, and T. Chevalier, "CIRDO: Smart companion for helping elderly to live at home for longer," *IRBM - Ingénierie et Recherche Biomédicale*, vol. 35, no. 2, pp. 101–108, Mar. 2014, 8 pages.

[23] J. F. Gemmeke, B. Ons, N. Tessema, H. V. hamme, J. van de Loo, G. D. Pauw, W. Daelemans, J. Huyghe, J. Derboven, L. Vuegen, B. V. D. Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces: an overview of the aladin project," in *Interspeech 2013*, 2013, pp. 2039–2043.

[24] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "homeService: Voice-enabled assistive technology in the home using cloud-based automatic speech recognition," in *4th Workshop on Speech and Language Processing for Assistive Technologies*, 2013.

[25] A. König, C. Crispim, A. Derreumaux, G. Bensadoun, P.-D. Petit, F. Bremond, R. David, F. Verhey, P. Aalten, and P. Robert, "Validation of an automatic video monitoring system for the detection of instrumental activities of daily living in dementia patients," *Journal of Alzheimer's Disease*, vol. 44, no. 2, pp. 675–685, 2015.

[26] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, pp. 2629–2634.

[27] M. Vacher, P. Chahuara, B. Lecouteux, D. Istrate, F. Portet, T. Joubert, M. E. A. Sehili, B. Meillon, N. Bonnefond, S. Fabre, C. Roux, and S. Caffiau, "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)*, Osaka, Japan, Jul. 2013, pp. 7298–7301.

[28] 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. 1–36, May 2015.

[29] P. Chahuara, F. Portet, and M. Vacher, "Making Context Aware Decision from Uncertain Information in a Smart Home: A Markov Logic Network Approach," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science, vol. 8309. Dublin, Ireland: Springer, 2013, pp. 78–93.

[30] M. Vacher, D. Istrate, and J. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees," in *Proc. 12th European Signal Processing Conference*, S. LTD, Ed., Vienna, Austria, sep. 2004, pp. 1171–1174.

[31] M. E. A. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, and J. Boudy, "Sound environment analysis in smart home," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science, vol. 7683. Pisa, Italy: Springer, 2012, pp. 208–223.

[32] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model—a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404 – 439, 2011, language and speech issues in the engineering of companionable dialogue systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088523081000063X

[33] 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 *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014, pp. 4499–4506.

[34] L. Zouari and G. Chollet, "Efficient gaussian mixture for speech recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4, 2006, pp. 294–297.

[35] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, M. Fapso, and J. Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Interspeech'05*, Lisboa, Portugal, 2005, pp. 633–636.

[36] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2338–2347, Nov 2011.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[38] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[39] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Estève, and C. Jacquin, "Automatic named identification of speakers using diarization and ASR systems," in *ICASSP 2009*, Taipei (Taiwan), 19-24 April 2009, pp. 4557–4560.

[40] J. Mauclair, S. Meignier, and Y. Estève, "Speaker diarization: About whom the speaker is talking ?" in *IEEE Odyssey*, San Juan, Puerto Rico, USA, 2006, pp. 1–6.

[41] Y. Estève, S. Meignier, and J. Mauclair, "Extracting true speaker identities from transcriptions," in *Interspeech 2007*, Antwerpen, Belgium, 2007, pp. 2601–2604.

[42] M. Vacher, B. Lecouteux, D. Istrate, T. Joubert, F. Portet, M. Sehili, and P. Chahuara, "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*, Grenoble, France, 2013, pp. 99–105.

[43] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, 2001.

[44] S. Novoselov, T. Pekhovsky, and K. Simonchik, "STC Speaker Recognition System for the NIST i-Vector Challenge," in *Odyssey 2014: The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014, pp. 231–240.

[45] J. Kahn, N. Audibert, S. Rossato, and J.-F. Bonastre, "Intra-speaker variability effects of Speaker Verification performance," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, Jun. 2010, p. n.c.

[46] G. R. Doddington, W. Liggett, A. F. Martin, M. A. Przybocki, and D. A. Reynolds, "Sheep, goats, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *The 5th International Conference on Spoken Language Processing, Incorporating The 7th Australian International Speech Science and Technology Conference, Sydney Convention Centre, Sydney, Australia, 30th November - 4th December 1998*, 1998. [Online]. Available: http://www.isca-speech.org/archive/icslp_1998/i98_0608.html