On Distant Speech Recognition for Home Automation

Michel Vacher, Benjamin Lecouteux, François Portet

On Distant Speech Recognition for Home Automation

Michel Vacher, Benjamin Lecouteux, and François Portet

CNRS, LIG, F-38000 Grenoble, France
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
Equipe GETALP
41 rue des Mathématiques, 38041 Grenoble Cedex9, France
{Michel.Vacher,Benjamin.Lecouteux,Francois.portet}@imag.fr

Abstract. In the framework of Ambient Assisted Living, home automation may be a solution for helping elderly people living alone at home. This study is part of the Sweet-Home project which aims at developing a new home automation system based on voice command to improve support and well-being of people in loss of autonomy. The goal of the study is vocal order recognition with a focus on two aspects: distance speech recognition and sentence spotting. Several ASR techniques were evaluated on a realistic corpus acquired in a 4-room flat equipped with microphones set in the ceiling. This *distant speech* French corpus was recorded with 21 speakers who acted scenarios of activities of daily living. Techniques acting at the decoding stage, such as our novel approach called Driven Decoding Algorithm (DDA), gave better speech recognition results than the baseline and other approaches. This solution which uses the two best SNR channels and *a priori* knowledge (voice commands and distress sentences) has demonstrated an increase in recognition rate without introducing false alarms. Generally speaking, a short overview allows then to outline the research challenges that speech technologies must take up for Ambient Assisted Living and Augmentative and Alternative Communication, and the current reseach avenues in this domain.

1 Introduction

Demographic change and ageing in developed countries are challenging the society effort in improving the well being of its elderly and frail inhabitants. The evolution of the Information and Communication Technologies led to the emergence of Smart Homes equipped with ambient intelligence technology which provides high man-machine interaction capacity [1]. However, the technical solutions implemented in such Smart Homes must suit the needs and capabilities of their users in the context of *Ambient Assisted Living*. Under some circumstances, classic tactile commands (e.g., the switch of the lamplight) may not be adapted to the aged population who have some difficulties in moving or seeing. Therefore, tactile commands can be complemented by speech based solutions that would provide voice command and would make it easier for the person to interact with her relatives or with professional carers (notably in

case of distress situations) [2]. Moreover, analysis of sounds emitted in a person's habitation may be useful for activity monitoring and context awareness.
The SWEET-HOME project was set up to integrate sound based technology within smart homes to provide natural interaction with the home automation system at any time and from anywhere in the house. As emphasized by Vacher et al. [3], major issues still need to be overcome. For instance, the presence of uncontrolled noise is a real obstacle for distant speech recognition and identification of voice commands in continuous audio recording conditions when the person is moving and acting in the flat. Indeed, it is not always possible to force the user to take up a position at a short distance and in front of a microphone when he has to manage a specific device, such as a remote control device. Therefore, some microphones are set in the ceiling to be available without any action of the user.
This paper presents preliminary results of speech recognition techniques evaluated on data recorded in a flat by several persons in a daily living context. A glossary is given in Section 2 in order to define all specific terms used in this chapter. The background, the state of the art and the challenges to tackle are given in Section 3. The data recording and the corpus are presented in Section 4. In Section 6, several techniques of multisource speech recognition are detailed and evaluated. Section 6.5 is devoted to word spotting needed to recognize voice commands in sentences. The chapter finishes with Section 7 which makes a review of the open problems with regard to the application of speech processing for Assistive Technologies and with Section 8 which emphasizes the future work and studies necessary to design a usable system in the real world.

2 Glossary

Activities of daily living (ADL) are, as defined by the medical community, the things we normally do in daily living, including any daily activity we perform for self-care (such as feeding ourselves, bathing, dressing, grooming), work, and leisure. Health professionals routinely refer to the ability or inability to perform ADLs as a measurement of the functional status of a person, particularly in regard to people with disabilities and the elderly. A well known scale for ADL was defined by Katz and Akporn [4].

Ambient Assisted Living (AAL) aims to help seniors to continue to manage their daily activities at home thanks to ICT solutions for active and healthy ageing.

Automatic Speech Recognition (ASR) is the translation of spoken words into text by an automatic analysis system.

Blind Source separation (BSS) is the separation of a set of source signals from a set of mixed signals, without the aid of additional information (or with very little information) about the source signals or the mixing process.

Driven Decoding Algorithm (DDA) is a method that allows to drive a primary system search by using the one-best hypotheses and the word posteriors gathered from a secondary system in order to improve the recognition performances.

Distant Speech Recognition is a particular case of Automatic Speech Recognition when the microphone is moved away from the mouth of the speaker. A broad variety of effects such as background noise, overlapping speech from other speakers, and reverberation are responsible of the high degradation of performances of the conventional ASR in this configuration.

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a Markov process with unobserved (hidden) states.

Home Automation is the residential extension of building automation. Home automation may include centralized control of lighting, appliances and other systems, to provide improved convenience, comfort, energy efficiency and security.

Home Automation Network is a network specially designed to ensure the link between sensors, actuators and services.

KNX (KoNneX) is a worldwide ISO standard (ISO/IEC 14543) for home and building control.

Maximum A Posteriori (MAP) estimator, as the maximum likelihood method, is a method that can be used to estimate a number of unknown parameters, such as parameters of a probability density, connected to a given sample. This method is related to maximum likelihood however, it differs in the ability to take into account a non-uniform a priori on the parameters to be estimated.

Maximum Likelihood Linear Regression (MLLR) is an adaptation technique that uses small amounts of data to train a linear transform which, in case of Gaussian distribution, warps the Gaussian means so as to maximize the likelihood of the data.

Recognizer Output Voting Error Reduction (ROVER) is based on a 'voting' or re-scoring process to reconcile differences in ASR system outputs. It is a post-recognition process which models the output generated by multiple ASR systems as independent knowledge sources that can be combined and used to generate an output with reduced error rate.

Smart Home is a house that is specially equipped with devices giving it the ability to anticipate the needs of their inhabitants while maintaining their safety and comfort.

Signal to Noise Ratio (SNR) it is a measure that compares the level of a desired signal to the level of a reference or to background noise: $SNR = \frac{P_{signal}}{P_{reference}}$ The signal and the noise are usually measured across the same impedance and the SNR is generally expressed in dB scale: $SNR_{dB} = 10.\log_{10}\left(\frac{P_{signal}}{P_{reference}}\right) = 20.\log_{10}\left(\frac{A_{signal}}{A_{reference}}\right)$, where P and A denote respectively the power and amplitude of signal or reference.

Wizard of Oz It is an interaction method in which the user is not informed that the reaction of a device is actually controlled by a human (the 'wizard'). This is a reference to the 1939 American musical fantasy film "The Wizard of Oz".

Word Error Rate (WER) is a common metric of the performance of a speech recognition or machine translation system. $WER = \frac{S+D+I}{N}$, where S is the number of substitutions, D the number of deletions, I the number of insertions, N the number of words in the reference.

Word Spotting is related to search and retrieval of a word in an audio stream.

3 Background and state of the art

As reported in section 1, Smart Homes have been designed with the aim of allowing seniors to keep control of their environment and to improve their autonomy. Despite the fact that audio technology has a great potential to become one of the major interaction modalities in Smart Home, this modality is seldom taken into consideration [5][6][7][8]. The most important reason is that audio technology as not reached a sufficient stage of maturity and that there is still some challenges to overcome [3]. The SWEET-HOME project presented in Section 3.1 aims at designing an audio analysis system running in real-time for voice commands recognition in a realistic home automation context. The state of the art and the challenges to tackle are developed in Section 3.2 while Section 3.3 focuses on keyword spotting.

3.1 The SWEET-HOME project

Main goals The SWEET-HOME project is a French national supported research project (`http: //sweet-home.imag.fr/`). It aims at designing a new smart home system by focusing on three main aspects: to provide assistance via *natural man-machine interaction* (voice and tactile command), to ease *social inclusion* and to provide *security reassurance* by detecting situations of distress. If these aims are achieved, then the person will be able to pilot his environment at any time in the most natural way possible [9].
Acceptance of the system is definitely a big issue in our approach therefore, a qualitative user evaluation was performed to assess the acceptance of vocal technology in smart homes [10] at the beginning of the project

and before the study presented in section 4. Height healthy persons between 71 and 88 years old, seven relatives (child, grand-child or friend) and three professional carers were questioned in co-discovery in a fully equipped smart home alternating between interview and Wizard of Oz periods. Important aspects of the project have been evaluated: voice command, communication with the outside world, domotic system interrupting a person's activity, and electronic agenda. In each case, the voice based solution was far better accepted than more intrusive solutions. Thus, in accordance with other user studies [11][12], audio technology seems to have a great potential to ease daily living for elderly and frail persons. To respect privacy, it must be emphasized that the adopted solution will analyse the audio information on the fly and is not designed to store the raw audio signal. Moreover, the speech recognizer must be made to recognize only a limited set of predefined sentences in order to prevents recognition of intimate conversations.

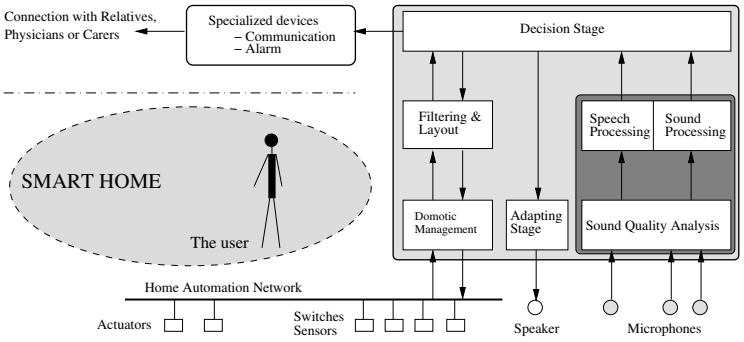

Fig. 1. The general organisation of the SWEET-HOME system

SWEET-HOME **technical framework** The SWEET-HOME system is depicted in Figure 1. The input of the system is composed of the information from the domotic system transmitted via a local network and information from the microphones transmitted through radio frequency channels. While the domotic system provides symbolic information, raw audio signals must be processed to extract information from speech and sound. This extraction is based on our experience in developing the AUDITHIS system [13], a real-time multi-threaded audio processing system for ubiquitous environments. The extracted information is analysed and either the system reacts to an order given by the user or the system acts pro-actively by modifying the environment without an order (e.g., turns off the light when nobody is in the room). Output of the system thus includes domotic orders, but also interaction with the user when a vocal order has not been understood for instance, or in case of alert messages (e.g., turn off the gas, remind the person of an appointment). The system

can also make it easier for the user to connect with her relative, physician or caregiver by using the e-lio[1] or Visage[2] systems. In order for the user to be in full control of the system and also in order to adapt to the users' preferences, three ways of commanding the system are possible: voice order, PDA or classic tactile interface (e.g., switch).
The project does not include the definition of new communication protocols between devices. Rather than building communication buses and purpose designed material from scratch, the project tries to make use of already standardised technologies and applications. As emphasized in [14], standards ensure compatibility between devices and ease the maintenance as well as orient the smart home design toward cheaper solutions. The interoperability of ubiquitous computing elements is a well known challenge to address [15]. Another example of this approach is that SWEET-HOME includes systems which are already specialised to handle the social inclusion part. We believe this strategy is the most realistic one given the large spectrum of skills that are required to build a complete smart home system.

3.2 Automatic Speech Recognition in Smart Homes

Automatic Speech Recognition systems (ASR) are especially good with close talking microphones (e.g., head-set), but the performances are significantly lower when the microphone is far from the mouth of the speaker such as in smart homes where microphones are often set in the ceiling. This deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise. All these problems are still to solve and should be taken into account in the home context.

Reverberation Distorted signals can be treated in ASR either at the acoustic model level or at the input (feature) level [16]. Deng et al [17] showed that feature adaptation methods provide better performances than those obtained with systems trained with data with the same distortion as the ones coming from the target environment (e.g., acoustic model learned with distorted data) for both stationary and non stationary noise conditions. Moreover, when the reverberation time is above 500 ms, ASR performances are not significantly improved when the acoustic models are trained on distorted data [18]. In the home involved in the study, the only glazed areas that are not on the same wall are right-angled, thus the reverberation is minimal. Given this and the small dimensions of the flat we can assume that the reverberation time stays below 500 ms. Therefore, only classic ASR techniques with adaptation using data recorded in the test environment will be considered in this study.

Background noise When the noise source perturbing the signal of interest is known, various noise removal techniques can be employed

[1] www.technosens.fr
[2] camera-contact.com

[19]. It is then possible to dedicate a microphone to record the noise source and to estimate the impulse response of the room acoustic in order to cancel the noise [20]. This impulse response can be estimated through Least Mean Square or Recursive Least Square methods. In a previous experiment, these methods showed promising results when the noise is composed of speech or classic music [21]. However, in case of unknown noise sources, such as washing machine or blender noise, Blind Source Separation (BSS) techniques seem more suited. The audio signals captured by the microphones are composed of a mixture of speech and noise sources. Independent Component Analysis is a subcategory of BSS which attempts to separate the different sources through their statistical properties (i.e., purely data driven). This method is particularly efficient for non-Gaussian signals (such as speech) and does not need to take into account the position of the emitter or of the microphones, but it assumes signal and noise to be linearly mixed, this hypothesis seems to be not suited in realistic recordings. Therefore, despite the important effort of the community, noise separation in realistic smart home condition remains an open challenge.

3.3 Word spotting

Spoken word detection has been extensively studied in the last decades especially in the context of spoken term detection in large speech databases and in continuous speech streams. Performances reported in the literature are good in clean conditions, especially with broadcast news data however, when experiences are undertaken in users' home conditions such as with noisy or spontaneous speech, performances decrease dramatically [22]. In [23], an Interactive Voice Response system was set up to support elderly people to deal with their medication. Over the 300 persons recruited, a third stopped the experiment because they complained about the system and only 38 persons completed the experiment.
In this study, some aspects of both spotting and Large Vocabulary Continuous Speech Recognition are considered. A Large Vocabulary Continuous Speech Recognition system was used in the approach to increase the recognition robustness. Language and acoustic models adaptation and multisource based recognition were investigated. Finally, we designed an original approach which integrates word matching directly inside the ASR system to improve the detection rate of domotic order, this will be described in section 6.5.

4 Recorded corpus and experimental framework

One experiment was conducted to acquire a multimodal corpus by recording individuals performing activities of daily living in a smart home. The speech part of the corpus, called the Sweet-Home *speech corpus*, is composed of utterances of domotic orders, distress calls and anodin sentences in French recorded using several microphones set in the ceiling of the smart home. This corpus was used to tune and to test a classic ASR system in different configurations. This section briefly introduces

the smart home, the SWEET-HOME *speech corpus*. The monosource ASR system is described in section 5.

4.1 Data acquisition in the smart home

The DOMUS smart home The SWEET-HOME speech corpus was acquired in realistic conditions, i.e., in a smart-home and in distant speech condition inside the DOMUS smart home. This smart home was designed and set up by the Multicom team of the Laboratory of Informatics of Grenoble to observe users' activities interacting with the ambient intelligence of the environment. Figure 2 shows the details of the flat. It is a thirty square meters suite flat including a bathroom, a kitchen, a bedroom and a study, all equipped with sensors and effectors. More than 150 sensors, actuators and information providers are managed in the flat. The flat is fully usable and can accommodate a dweller for several days so that it is possible to act on the sensory ambiance, depending on the context and the user's habits. The technical architecture of DOMUS is based on the KNX bus system (KoNneX), a worldwide ISO standard (ISO/IEC 14543) for home and building control. The flat has also been equipped with 7 radio microphones for the need of the SWEET-HOME project; the microphones are set into the ceiling (2 per room except for the bathroom). Audio data can be recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card [13]. The sample rate is 16kHz and the bandwith 8kHz. It must be noticed that the distance between the speaker and the closest microphone is about 2 meters when he is standing and about 3 meters when he is sitting.

Corpus recording 21 persons (including 7 women) participated to a 2-phase experiment to record, among other data, speech corpus in the DOMUS smart home. To make sure that the audio data acquired would be as close as possible to real daily living sounds, the participants performed several daily living activities. Each experimental session lasted about 2 hours. The average age of the participants was 38.5 ± 13 years (22-63, min-max). No instruction was given to any participant about how

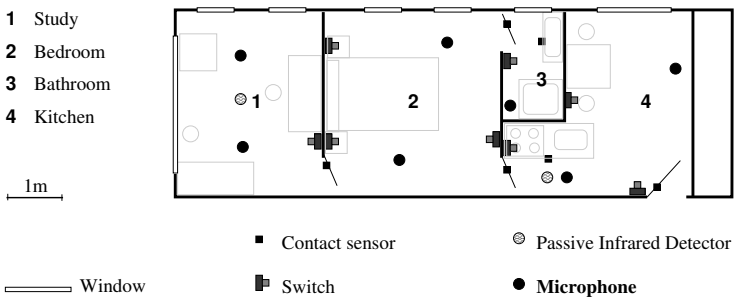

Fig. 2. The DOMUS Smart Home and the position of the sensors

they should speak and in which direction. Consequently, no participant emitted sentences directing their voice to a particular microphone.
A visit, before the first phase of the experiment, was organized to make the participants accustomed to the home in order to smoothly perform the experiment. During this first phase, participants uttered forty predefined French casual sentences on the phone such as "Allo" (*Hello*), "J'ai eu du mal à dormir" (*I slept badly*) but were also free to utter any sentence they wanted (some did speak to themselves aloud). Then, the first phase consisted in following a scenario of activities without condition on the time spent and the manner of achieving them (having a breakfast, listening to music, get some sleep, clean up the flat using the vacuum, etc.). Note that the microphone of the telephone was not recorded, only the 7 microphones set on the ceiling were used.
The second phase consisted in reading aloud a list of 44 sentences:

- 9 distress sentences such as "A l'aide" (*Help*), "Appelez un docteur" (*call a doctor*);
- 3 orders such as "Allumez la lumière" (*turn on the light*);
- 32 colloquial sentences such as "Le café est très chaud" (*The coffee is hot*).

This list was read in 3 rooms (study, bedroom, and kitchen) under three conditions: no background noise, vacuum on or radio on. 396 sentences were recorded but only those in the clean condition were used in this paper, the noisy condition records having been designed for other experiments.

4.2 The Sweet-Home French speech corpus

Only the sentences uttered in the study during the phone conversation of the phase 1 were considered. For the phase 2 record, only the sentences uttered in the kitchen without additional noise (vacuum or radio) were considered. Each speaker did not follow strictly the instructions given at the beginning of the experiment, therefore this corpus was indexed manually. Some hesitations and word repetitions occurred along the records. Moreover, when two sentences were uttered without a sufficient silence between them, they were considered as one sentence. A complete description of the corpus according to each speaker is given in Table 1. The Sweet-Home speech corpus is made of 862 sentences uttered by 21 persons in the first phase, 917 sentences in the second phase; it lasts for each channel 38 minutes 46s in the case of the first phase, and 40 minutes 27 s in the case of the second phase. The SNR (Signal-to-Noise Ratio) is an important parameter which was used for the combination of several sources. For Phase 1 (when the speaker was in the study) mean SNR was 21.8 dB/20.0 dB (channels 6 and 7), for Phase 2 (when the speaker was in the bedroom) mean SNR was 22.1 dB/22.1 dB (channels 4 and 5).
The databases recorded in the course of the Sweet-Home project are devoted to voice controlled home automation, they will be distributed for an academic and research use only [24].

Table 1. SWEET-HOME speech corpus description

Spkr. ID	**Phase 1** Duration (s) Channel 6 or 7	SNR mean (dB) Channel 6	SNR mean (dB) Channel 7	**Phase 2** Duration (s) Channel 4 or 5	SNR mean (dB) Channel 4	SNR mean (dB) Channel 5
1	145.78	23.5	22.1	96.66	24.7	26.2
2	119.36	22.6	21.0	110.42	21.2	22.0
3	112.08	14.8	12.2	119.76	15.9	16.7
4	141.32	16.5	16.5	119.04	22.1	24.0
5	159.32	29.7	26.8	122.21	26.8	28.6
6	122.10	17.7	16.1	108.61	19.7	18.7
7	110.90	19.0	17.5	116.00	20.7	21.2
8	114.54	20.3	19.0	114.64	18.9	20.6
9	121.58	26.8	24.7	135.36	24.5	25.3
10	77.50	20.3	18.0	104.54	23.4	18.8
11	106.52	20.2	21.0	105.76	20.6	23.9
12	90.48	24.5	21.1	108.44	25.1	24.3
13	96.46	26.2	19.9	116.52	17.3	13.2
14	97.74	17.7	17.7	113.40	18.5	15.3
15	96.48	22.6	21.4	101.98	25.0	26.9
16	96.86	21.4	17.6	106.72	18.2	10.7
17	111.08	21.7	20.0	144.46	28.3	24.6
18	169.14	20.0	19.0	124.52	23.0	24.2
19	146.98	25.1	23.4	125.58	24.4	22.4
20	89.80	27.5	24.8	120.60	29.0	27.4
21	99.48	19.5	19.2	109.56	17.4	14.4
Average	115.50	21.8	20.0	115.47	22.1	20.4

5 Monosource ASR techniques

The architecture of an ASR is described by Figure 3. A first stage is the audio interface in charge of acoustical feature extraction in consecutive frames. The next 3 stages working together are:

- the phoneme recognizer stage;
- the word recognition stage constructing the graph of phonemes; and
- the sentence recognition stage constructing the graph of words.

The data associated with these stages are respectively the acoustic models, the phonetic dictionary and the language models. The output of the recognizer is made of the best hypothesis lattices.

5.1 The Speeral ASR system

The ASR system used in the study is Speeral [25]. The LIA (Laboratoire d'Informatique d'Avignon) speech recognition engine relies on an A^* decoder with HMM-based context-dependent acoustic models and trigram

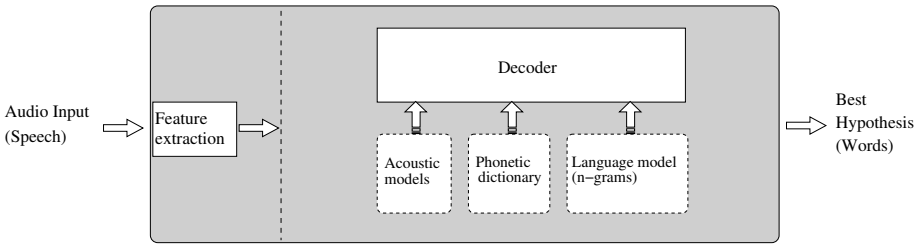

Fig. 3. General organisation of an ASR

language models. HMMs are classic three-state left-right models while state tying is achieved by using decision trees. The acoustic features, for each 30ms-length frame with 20ms overlay (10ms-time shift), were composed of 12 Perceptual Linear Predictive coefficients, the energy, and the first and second order derivatives of these 13 parameters, this represent in total 39 parameters. The acoustic models were trained on about 80 hours of annotated French speech. If the participants were elderly people, the use of adapted data would be required [26], but this was not the case for this study. Given the targeted application of Sweet-Home the computation time should not be a breach of real-time use. Thus, the $1\times$RT Speeral configuration was used. This this configuration, by using a strict pruning scheme, the time spent by the system to decode one hour of speech signal is real-time.

Language models Two language models were built: the generic and the specialized models. The *specialized* language model was estimated from the sentences that the 21 participants had to read during the experiment (domotic orders, casual phrases, etc.). The *generic* language model was estimated on about 1000M of words from the French newspapers *Le Monde* and *Gigaword*.

5.2 Baseline system

In order to propose a **baseline system**, the adaptation of both acoustic and language models were tested. Then, to improve the robustness of the recognition, multi-streams ASR was tested. Finally, a new variant of a driven decoding algorithm was used in order to take into account *a-priori* information and several audio channels for each speaker.
The phase 1 of the corpus was used for development and acoustic model adaptation to the speaker while the phase 2 was used for performances estimation. Results obtained on the phase 2 of the corpus were analysed using two measures: the Word Error Rate (WER) and the Classification Error Rate (CER). The WER is a good measure of the robustness, while the CER corresponds to the main goal of our research (i.e., detection of *predefined* sentences).

Acoustic models adaptation: MAP versus MLLR Acoustic models were adapted for each speaker by using two methods: Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR) by using data of the first phase. These data were perfectly annotated, allowing to perform correct targeted speaker adaptation.

The Maximum Likelihood Linear Regression (MLLR) is used when a limited amount of data per class is available. MLLR is an adaptation technique that uses small amounts of data to train a linear transform which warps the Gaussian means so as to maximize the likelihood of the data : acoustically close classes are grouped and transformed together. In the case of the Maximum a posteriori approach (MAP), initial models are used as informative priors for the adaptation.

Table 2 shows different results with and without acoustic models adaptation. Results are presented for the two best streams (high SNR). Experiments were carried out with the generic language model (GLM) lightly interpolated with *predefined* sentences (PS) presented in the next section. Without acoustic adaptation, the best average WER is about 36%. The results show that MAP is not very performing in this case. With MAP, the WER about 27%. The best average WER is about 18% with MLLR adaptation, which is the best choice for sparse and noisy data whatever the channel.
Two aspects explain the MAP performance:

- The noisy environment is not adapted to MAP adaptation [27].
- The lack of parameter tying in the standard MAP algorithm implies that the adaptation is not robust.

Linguistic variability Large vocabulary model languages such as the *generic* language model, are known to perform poorly on specific tasks because of the large number of equi-probable hypotheses. Better recognition can be obtained by reducing the overall linguistic space by estimating a language model on the expected sentences such as with the *specialized* language model. However, such a language model would be probably too specific when the speaker deviates from the original transcript. To benefit from the two language models, we propose a linear interpolation scheme where specific weights are tested on specialized and generic language models. The reduction of the linguistic variability thanks to the contribution of known *predefined* sentences is explored. Therefore, we interpolated the specialized model with the generic large vocabulary language model.
Two schemes of linear interpolation were considered: in the first one, the *generic* model had a strong weight while in the second one, the impact of the *generic* model was low. The ASRs were assessed after MLLR adaptation using the data of phase 1 of the corpus. Table 2 presents the WER with the generic language model (Baseline). As expected, the baseline language model obtained poor results: about 74%. Without reliable information, the ASR system, in noisy, speaker independent and large vocabulary condition is unable to perform good recognition.

Table 2. Average WER according to different configurations by using monosource techniques

Method	WER stream 1 channel 4 (%)	WER stream 2 channel 5 (%)
Generic LM	75.3	73.4
Interpolated LM	38.8	35.0
Interpolated LM with MAP adaptation	28.5	25.9
Interpolated LM with MLLR adaptation	18.6	18.0
Specialized LM with MLLR adaptation	19.2	19.0

With the specialised language model the system is able to detect more *predefined* sentences. However, when the speaker deviates from the scenario, the language model is unable to find the correct uttered sentence. The specialised language model was thus too specific.

Finally, a light (10%) interpolated language model led to the best results. This model combined the generic language model (with a 10% weight) and the specialised model (with 90% weight). These results show that a decoding based on a language model mainly learnt from the *predefined* sentences improves significantly the WER. The best WER is obtained when a generic language model is also considered: when the speaker deviates, the generic language model makes it possible to correctly recognise the pronounced sentences.

5.3 Conclusion about monosource ASR

Speeral ASR system was evaluated taking into account realistic distant-speech conditions and in the context of a home automation application (voice command). The system had to perform ASR with several constraints and challenges. Indeed, the noisy, distant-speech conditions, speaker independent recognition, continuous analysis and real-time aspects, the analysis system must operate in more difficult conditions than with the classic head-set one. Therefore, it is clear that obtained results are insufficient and must be improved, multichannel analysis is an avenue worth exploring.

The application conditions also make it possible for the ASR system to benefit from multiple audio channels, from a reduced vocabulary and from the hypothesis that only one speaker should utter voice commands. Lightly interpolated language model and a MLLR acoustic adaptation did improve significantly the ASR system performance. In the next section, we propose several techniques based on this baseline in order to perform multisource ASR.

6 Techniques for multisource speech recognition and sentence detection

Multisource ASR can improve the recognition performances thanks to information extracted in more than one channel. The ROVER method presented in Section 6.1 analyses the outputs of ASR performed on all channels separately. In the DDA method presented in Section 6.2, the information of one channel is used to guide the analysis on another channel. We also present an improved DDA method in Section 6.3 were *a priori* information about the task is taken into account.

6.1 ROVER

At the ASR combination level, a ROVER [28] was applied. ROVER is expected to improve the recognition results by providing the best agreement between the most reliable sources. It combines systems output into a single word transition network. Then, each branching point is evaluated with a vote scheme. The words with the best score are selected (number of votes weighted by confidence measures). However, this approach necessitates high computational resources when several sources need to be combined and real time is needed (in our case, 7 ASR systems must operate concurrently).
A baseline ROVER was tested using all available channels without *a priori* knowledge. In a second step, an *a priori* confidence measure based on the SNR was used: for each decoded segment s_i from the i^{th} ASR system, the associated confidence score $\phi(s_i)$ was computed according to equation 1 where $R()$ is the function computing the SNR of a segment and s_i is the segment generated by the i^{th} ASR system:

$$\phi(s_i) = 2^{R(s_i)} / \sum_{j=1}^{7} 2^{R(s_j)} \tag{1}$$

For each annotated sentence a silence period I_{sil} at the beginning and the end is taken around the speech signal period I_{speech}. The SNR is thus evaluated through the function $R()$ according to Equation 2.

$$R(S) = 10 * log(\frac{\sum_{n \in I_{speech}} S[n]^2}{|I_{speech}|} / \frac{\sum_{n \in I_{sil}} S[n]^2}{|I_{sil}|}) \tag{2}$$

Finally, a ROVER using only the two best channels overall was tested in order to check whether other channels contain redundant information and whether good results can be reached with low computational cost.
The ROVER combination led to great improvements. The results show that the ROVER made ASR more robust with an average WER of 13.0%. This aspect shows the complementarity of the streams. However, the ROVER stage increased the computation time proportionally to the number of ASR systems used. Given that the objective of the project is to build a real-time and affordable solution, computational resources are limited. Moreover, ROVER combination for two streams reduces the problem to picking the word with the highest confidence

when two systems disagree. Thus, when the recogniser confidence scores are not reliable, the ROVER between two streams does not perform well and the final performance is likely to be similar to a single system. Thus, we propose in the next section a method allowing low-cost computations with only two streams, based on the Driven Decoding Algorithm. In the following, ROVER results are used as baseline.

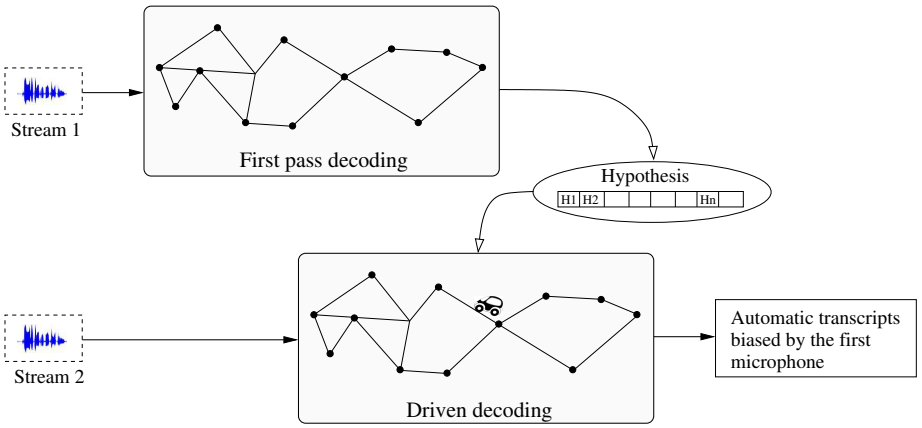

Fig. 4. Driven Decoding Algorithm used with two streams: The first stream drives the second stream

6.2 Driven Decoding Algorithm

The Driven Decoding Algorithm (DDA) [29, 30] is able to simultaneously align and correct the imperfect ASR outputs [31]. DDA has been implemented within Speeral: The ASR generates assumptions as it walks the phoneme lattice. For each new step, the current assumption is aligned with the approximated hypothesis. Then, a matching score α is computed and integrated within the language model:

$$\tilde{P}(w_i|w_{i-1}, w_{i-2}) = P^{1-\alpha}(w_i|w_{i-1}, w_{i-2}) \tag{3}$$

where $\tilde{P}(w_i|w_{i-1}, w_{i-2})$ is the updated trigram probability of the word w_i given the history w_{i-2}, w_{i-3}, and $P(w_i|w_{i-1}, w_{i-2})$ is the initial probability of the trigram. When the trigram is aligned, α is at a maximum and decreases according to the misalignments of the history (values of α must be determined empirically using a development corpus).
In the Domus smart home, uttered sentences were recorded using two microphones per room. Thus, two microphones can be used as input to DDA in order to increase the robustness of the ASR systems as presented in Figure 4. We propose to use a variant of the DDA where the output of

the first microphone is used to drive the output of the second one. This approach presents two main benefits:

- The second ASR system speed is boosted by the approximated transcript (only 0.1×RT)
- While a ROVER does not allows to combine efficiently two systems without confidence scores, DDA combines easily the information

The Figure 4 explains the Driven Decoding solution: the first Speeral pass on the stream 1 is used to drive a second pass on the stream 2, allowing to combine the information of the two streams.
Results using the 2-stream DDA are presented in Table 3. In most cases, DDA generated hypotheses that led either to the average WER of the two initial streams or to better WER. The average WER is 11.4%. We propose to extend this approach in the next section by driving the ASR system by *a priori* sentences selected on the first stream.

6.3 Two level DDA

In the previous approach, the first stream of decoding was used to drive the second one: DDA aims to refine the decoding achieved during the first stream decoding. Word spotting using ASR systems is known to be focused on accuracy, since the prior probability of having the targeted terms in a transcription is low. On the other hand, transcription

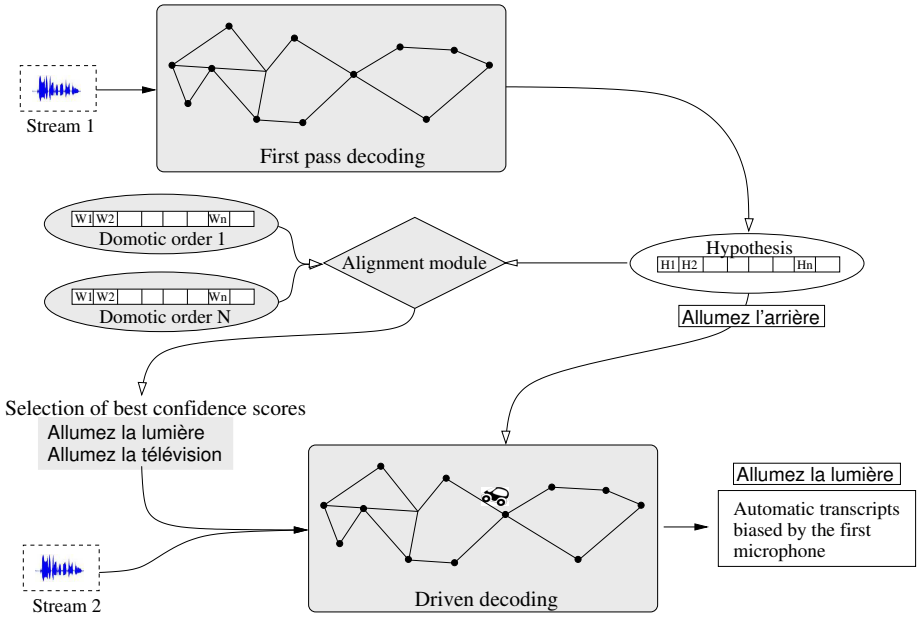

Fig. 5. Driven Decoding Algorithm used with two streams and *a priori* sentences: The first stream drives the second stream according to a refine selection of spotted sentences

Table 3. ASR system recognition WER by using multisource techniques

Method	WER $(\%)^{\pm SD}$
Baseline	$18.3^{\pm 12.1}$
Oracle Baseline	$17.7^{\pm 10.3}$
ROVER Full	$20.6^{\pm 8.5}$
ROVER 2 channels + SNR	$13.0^{\pm 6.6}$
ROVER +SNR	$\mathbf{12.2}^{\pm 6.1}$
DDA +SNR	$11.4^{\pm 5.6}$
DDA 2 levels + SNR	$\mathbf{8.8}^{\pm 3.7}$

errors may introduce mistakes and lead to misses of correct utterances, especially on large requests: the longer the searched term, the higher the probability of encountering an erroneous word. In order to limit this risk, we introduced a two-level DDA: speech segments of the first pass are projected in $3 - best$ spotted sentences and injected via DDA into the ASR system for the second decoding pass. The first decoding pass allows to generate hypotheses. By using the edit distance explained in 6.5, closed spotted sentences are selected and used as input for the fast second pass as presented in Figure 5. In this configuration, the first pass is used to select some sentences used to drive the second pass. In the Figure 5, the first system outputs " Allumer la lumière" (*Turn on the light*). The edit distance allows to find two close sentences: "Allumez la lumière" and "Allumez la télévision" (*Turn on the TV*). These sentences drive the second pass and allows one to find the correct output "Allumez la lumière".
Results using this approach are showed in Table 3. According to the WER, this approach improved significantly the ASR system quality, by taking advantage of the *a priori* information assessed by the *predefined* spotted sentences. WER is improved significantly for all speakers: the mean WER is 8.8%. By using the two streams available the ASR system is able to combine them efficiently. The best results are obtained with the two level approach were the ASR system is driven by both the first stream and the potential spotted sentences. The next section investigates the impact of each previous proposed method on the detection of pronounced sentences.

6.4 Multisource speech recognition: results

For each approach, the presented results are the average over the 21 speakers (plus standard deviation for the WER). For the sake of comparison, results of a baseline and an oracle baseline systems are provided. The baseline system outputs the best decoding amongst 7 ASR systems according to the highest SNR. The oracle baseline is computed by selecting the best WER for each speaker. The best results are achieved with

DDA because the search for the best hypothesis in the lattice uses data from several channels and has more information than when decoding for each channel.

6.5 Detection of predefined sentences

In order to spot sentences into automatic transcripts T of size m, each sentence of size n from *predefined* sentences H was aligned with T by using a Dynamic Time Warping (DTW) algorithm at the letter level [32]. Sequences were aligned by constructing an n-by-m matrix where the (i^{th}, j^{th}) element of the matrix contained the distance between the two words T_i and H_j using the distance function defined below.

$$\begin{array}{l} d(T_i, H_j) = 0 \text{ if } T_i = H_j \\ d(T_i, H_j) = 3 \text{ in the insertion cases} \\ d(T_i, H_j) = 3 \text{ in the deletion cases} \\ d(T_i, H_j) = 6 \text{ in the substitution cases} \end{array} \quad (4)$$

The deletion, insertion and substitution costs were computed empirically. The cumulative distance $\gamma(i, j)$ between H_j and T_i is computed as:

$$\gamma(i, j) = d(T_i, H_j) + min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (5)$$

Each *predefined* sentence is aligned and associated with an alignment score: the percentage of well aligned symbols (here letters). The sentence with the best score is then selected as best hypothesis.
This approach takes into account some recognition errors such as word declination or light variations (téléviseur, télévision etc.). Moreover, miss-decoded word are often orthographically close from the good one (due to the close pronunciation).
To test the detection of *a-priori* pronounced sentences, such as domotic orders (e.g., "allume la lumière"), the detection methods were applied in the following ASR configurations:

- Baseline: Speeral system with acoustic and language model adaptation.
- ROVER: Consensus vote between all streams.
- DDA1: DDA drived with the first stream.
- DDA2: DDA drived by the first stream and the spotted sentences.

The three systems based on ROVER and DDA gave the best performances, with respectively 88.2%, 87,4% and 92.5% of correct classifications while the baseline system obtains 85% of correct classification. It can be observed that the 2-level DDA based ASR system was able to detect more spotted sentences with less computational time and with more accuracy than the ROVER based one.

Sentence detection: results In all best-configurations, *predefined* sentence recognition showed a good accuracy: the baseline recognition gave 85%. It can be observed that in other configurations the spotting task correlated well with the WER. Thereby, ROVER and the two DDA configurations led to a significant improvement over the baseline. The best configuration based on the two-level DDA gave 92.5% of correct classifications.

6.6 Discussion and future works

The goal of this study is to provide a path for vocal command recognition improvement with a focus on two aspects: distance speech recognition and sentence spotting. A distant speech French corpus was recorded with 21 speakers playing scenarios of activities of daily living in a real flat, this corpus is made of colloquial sentences, vocal commands and distress sentences. This realistic corpus was acquired in a 4-room flat equipped with microphones set in the ceiling thanks to 21 speakers. Several ASR techniques were evaluated, such as our novel approach called Driven Decoding Algorithm (DDA). They gave better results than the baseline and other approaches. Indeed, they analyse the signal on the two best SNR channels and the use of a priori knowledge (specified vocal commands and distress sentences) increases the recognition rate in the case of true positive sentences and doesnt introduce false positive.

Evaluation in real conditions The technology developed in this study was then tested thanks to two other experiments in an Ambient Assisted Living context at the end of the Sweet-Home project. These experiments involved 16 non-aged participants for the first one and 11 aged or visually impaired people for the second one [33]. Each participant followed a scenario including various situations and activities of the daily life. The objective of these experiments was to evaluate the use of voice command for home automation in distant speech conditions, in real-time and in context aware conditions [34]. Unfortunately, we were not able to integrate the DDA method in time in the real-time analysis software PATSH before the beginning of these experiments. Therefore, the performance of the system was still low, the Home Automation Command Error Rate was about 38% [33], but the results showed there is room for improvement. But, although the participants had to repeat, sometimes up to three times, the voice command, they were overall very excited about commanding their own home by voice. These results highlight the interest of the methods discussed above and especially DDA2 that chooses among available channels those that have the best SNR in order to refine the data analysis. One of the biggest problems were the response time which was unsatisfactory (for 6 participants out of 16) and the mis-understanding of the system which implied to repeat the order (8/16). These technical limitations were reduced when we improved the ASR memory management and reduced the search space. After this improvement, only one participant with special needs complained about the response time.

Interest of the recorded corpus During these experiments, all data were recorded. This acquired corpus was used to evaluate the performance of the audio analysis methods presented in this chapter. It constitutes a precious resource for future work. Indeed, one of the main problems that impede researches in this domain is the need for a large amount of annotated data (for analysis, machine learning and benchmark). It is quite obvious that the acquisition of such datasets is highly

expensive both in terms of material and of human resources. For instance, in the experiment presented in section 4, the acquisition and the annotation of the 33-hours corpus costed approximatively 70k€.
Therefore, the Sweet-Home multimodal corpus is a dataset recorded in realistic conditions in Domus, the fully equipped Smart Home with microphones and home automation sensors presented in section 4.1 will be available for the research community [24]. This corpus was recorded thanks to participants which performed Activities of Daily living (ADL). This corpus is made of a multimodal subset, a French home automation speech subset recorded in Distant Speech conditions, and two interaction subsets, the first one being recorded by 16 persons without disabilities and the second one by 6 seniors and 5 visually impaired people. This corpus was used in studies related to ADL recognition, context aware interaction and distant speech recognition applied to home automation controlled through voice.

Future projects Our future project aims to develop a system capable of operating under the conditions encountered in an apartment. For this we must firstly integrate BSS techniques to reduce the noise present in the everyday life context and secondly improve the DDA2 method to detect and recognize the voice commands as well as distress calls.

7 Application of speech processing for Assistive Technologies

The applications of speech processing may present a greet benefit for smart homes and Ambient Assisted Living (see Section 7.1) but Augmentative and Alternative Communication (AAC) retains involvement from a broad community of researchers (see Section 7.2).

7.1 Smart home and AAL

Anticipating and responding to the needs of persons with loss of autonomy with ICT is known as Ambient Assisted Living (AAL). ICT can contribute to the prevention and / or compensation of impairments and disabilities, to improve the quality of life, safety, communication and social inclusion of end users. They must relieve the isolation and caregiver burden. They also participate in the modernization of health and social services by facilitating home or institutional organization of professional care, their implementation, their tolerance and performance [35]. In this domain, the development of smart homes is seen as a promising way of achieving in-home daily assistance [1]. Health Smart Home has been designed to provide daily living support to compensate some disabilities (e.g., memory help), to provide training (e.g., guided muscular exercise) or to detect potentially harmful situations (e.g., fall, gas not turned off). Basically, a health smart home contains sensors used to monitor the activity of the inhabitant. Sensor data are analyzed to detect the current situation and to execute the appropriate feedback or assistance

A rising number of studies about audio technology in smart home were conducted. This includes speech recognition [36][37][38][39], sound recognition [3][40][41], speech synthesis [42] or dialogue [7][8][43][11]. These systems are either embedded into the home automation system or in a smart companion (mobile or not) or both as in Companions [44] or CompanionAble [41] projects.

However, given the diverse profiles of the users (e.g., low/high technical skill, disabilities, etc.), complex interfaces should be avoided. Nowadays, one of the best interfaces, is the VoiceUser Interface (VUI), whose technology has reached a stage of maturity and that provides interaction using natural language so that the user does not have to learn complex computing procedures [10]. Moreover, it is well adapted to people with reduced mobility and to some emergency situations (hand free and distant interaction). Indeed, a home automation system based on voice command will be able to improve support and well-being of people in loss of autonomy. But, despite the interest presented by sound analysis techniques, the use of ASR for voice command for home automation in a real environment is still an open challenge.

Voice-User Interface in domestic environment has recently gained interest in the speech processing community as exemplified by the rising number of smart home projects that considers Automatic Speech Recognition (ASR) in their design [45][46][47][5][8][6][48][37][38][39][49]. However, though VUIs are frequently employed in close domains (e.g., smart phone) there are still important challenges to overcome [3]. Indeed, the task imposes several constraints to the speech technology:

- distant speech conditions [16],
- hand free interaction,
- adaptation to potential users (elderly),
- affordable by people who can have low resources,
- noise conditions in the home,
- real-time,
- respect of privacy.

In recent years, the research community shows an increased interest with regards to the analysis of the speech signal in noisy conditions like the organizing of Challenges *CHiME* shows. The first CHiME Challenge held in 2011 was the first concerted evaluation of ASR systems in a real-world domestic environment involving both reverberation and highly dynamic background noise made up of multiple sound source [50]. The second CHiME Challenge in 2013 was supported by the IEEE AASP, MLSP and SL Technical Committees [51]. The configuration considered by this Challenge was that of speech from a single target speaker being binaurally recorded in a domestic environment involving multisource background noise. These challenges reported here are still no close enough to real conditions and future editions of the challenge will attempt to move closer to realistic conditions.

Ageing has effects on the voice and movement of the person and thereby, aged voice is characterized by some specific features such as imprecise production of consonants, tremors, hesitations and slower articulation [52]. Some studies have shown age-related degeneration with atrophy of vocal cords, calcification of laryngeal cartilages, and changes in muscles

of larynx [53][54]. For there reason, some authors highlight that ASR performance decreases with elderly voice. This phenomenon has been observed in the case of English, European Portuguese, Japanese and French [55][56][57][26]. Vipperla et al [58] made a very useful and interesting longitudinal study by using records of defence speech delivered in the Supreme Court of the United States over a decade by the same judges. This study showed that an adaptation to each speaker can get closer to the scores of non-aged speakers but this implies that the ASR must be adapted to each speaker. Nevertheless, some authors established that many other effects can also be responsible for ASR performance degradation such as decline in cognitive and perceptual abilities [59][60]. Moreover, since smart home systems for AAL often concern distress situations, it is unclear whether distress voice will challenge the applicability of these system. Speech signal contains linguistic information but it may be influenced by the health, the social status and the emotional state [61][62]. Recent studies suggests that ASR performance decreases in case of emotional speech [63][64], however it is still an under-researched area. In their study, Vlasenko et al [63] demonstrated that acoustic models trained on read speech samples and adapted to acted emotional speech could provide better performance of spontaneous emotional speech recognition.
Moreover, such technology must be validated in real smart homes and with potential users. At this time, validation studies in such realistic conditions are rare [33]. In the same way, there are few user studies reported in the literature and related to speech technology application [10], they are generally related to ICT [65].

7.2 Assistive technologies

The field of Augmentative and Alternative Communication (AAC) is multidisciplinary and vast, its focus is to develop methods and technologies to aid communication for people with complex communications needs [66]. Potential users are elderly and all people who may acquire a disability or have a degenerative disability which affects communication, this disability can result from both motor and cognitive impairments (i.e., paralysis, hearing or visual impairment, brain injury, Alzheimer...).
Speech and language processing play a major role to improve function for people with communication facilities [67]. This is highlighted by the publication of special issues of journals and by the regular organisation of workshops and conferences on this topic. In 2009, the third issue of the ACM Transactions on Accessible Computing was devoted to AAC (Volume 1, Issue 3). In 2011, the relationship between assistive technology and computational linguistics was formalized with the formation of an ACL Special Interest Group on Speech and Language Processing for Assitive Technology (SIG-SLPAT[3]) which gained SIG status from the International Speech Communication Association (ISCA). The last workshops SIG-SLPAT bringing together Computational Linguistics, Speech Processing and Assistive Technologies took place in Montreal, Quebec

[3] http://www.slpat.org/

(2012), in Grenoble, France (2013) and in Baltimore, U.S. (2014). In the same way, a special session of Interspeech[4] "Speech technologies for Ambient Assisted Living" is organized in 2014. This special session aims at bringing together researchers in speech and audio technologies with people from the ambient assisted living and assistive technologies communities to meet and foster awareness between members of either community, discuss problems, techniques and datasets, and perhaps initiate common projects.
Regarding speech recognition, the most important challenges are related to the recognition of speech uttered by elderly, dysarthric or cognitively impaired speakers.

8 Future outlook

Future challenges have been outlined in the previous section 7. These challenges are essentially related to scientific and technological problems to solve, but the human aspect must not be neglected.

8.1 Scientific and technical challenges

In real home environment the audio signal is often perturbed by various and undetermined noises (e.g., devices, TV, music, roadwork...). But this also shows us the challenges to obtain a usable system that will not be set-up in lab conditions but in various and noisy ones. Of course, in the future, smart homes could be designed specifically to limit these effects but the current smart home development cannot be successful if we are not able to handle these issues when equipping old-fashioned or poorly insulated home. Finally, one of the most difficult problems is the blind source separation. Some techniques developed in other areas of signal processing may be considered to analyze speech captured with far-field sensors and to develop a Distant Speech Recogniser (DSR) such as blind source separation, independent component analysis (ICA), beam-forming and channel selection.
Two main categories of audio analysis are generally targeted: daily living sounds and speech. These categories represent completely different semantic information and the techniques involved for the processing of these two kinds of signal are quite distinct. However, the distinction can be seen as artificial and there is a high confusion between speech and sounds with overlapped spectrum. For instance, one problem is to know whether scream or sigh must be classified as speech or sound.
Moreover, the system must react as quickly as possible to a vocal order. For example, if the user says "Nestor allume la lumière" (Nestor turn on the light), the sentence duration is about 1s, and the processing time last generally between 1.5 and 2s. This duration seems low but this is not true in real conditions when the user in the obscurity is waiting for the light. Thus, optimisation are needed to obtain fast recognizers.

[4] http://www.interspeech2014.org/

8.2 Human aspect

One of the main challenges to overcome for successful integration of VUI in AAL, is the adaptation of the system to the elderly users. Indeed, the ageing process is characterised by a decay of the main bio-physiological functions, affecting the social role and the integration of the ageing person in the society. Overall elderly people will be less inclined to adapt to a technology and its limitation (e.g., the constraint to pronounce words in a certain way) than younger adults and will present a very diverse set of profiles that make this population very difficult to design for.
For the elderly, there is a balance between the benefit of a monitoring through sensors and the correspondent intrusion into privacy. The system has to be protected against intrusion and has to make sure that the information reaches only the right people or can not go out of the smart home.
This is the most important aspect because if the system is not accepted by its potential users, it will never be used in practice.

Acknowledgments. This work is part of the Sweet-Home project supported by the French National Research Agency (Agence Nationale de la Recherche / ANR-09-VERS-011).

References

1. Chan, M., Estève, D., Escriba, C., Campo, E.: A review of smart homes- present state and future challenges. Computer Methods and Programs in Biomedicine **91**(1) (2008) 55–81
2. Vacher, M., Portet, F., Rossato, S., Aman, F., Golanski, C., Dugheanu, R.: Speech-based interaction in an AAL context. Gerontechnology **11**(2) (jul 2012) 310
3. Vacher, M., Portet, F., Fleury, A., Noury, N.: Development of Audio Sensing Technology for Ambient Assisted Living: Applications and Challenges. International Journal of E-Health and Medical Communications **2**(1) (2011) 35–54
4. Katz, S., Akpom, C.: A measure of primary sociobiological functions. Journal of Health Services **6**(3) (1976) 493508
5. Badii, A., Boudy, J.: 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) 18–20
6. Filho, G., Moir, T.: From science fiction to science fact: a smart-house interface using speech technology and a photorealistic avatar. International Journal of Computer Applications in Technology **39**(8) (2010) 32–39
7. Gödde, F., Möller, S., Engelbrecht, K.P., Kühnel, C., Schleicher, R., Naumann, A., Wolters, M.: Study of a speech-based smart home system with older users. In: International Workshop on Intelligent User Interfaces for Ambient Assisted Living. (2008) 17–22

8. Hamill, M., Young, V., Boger, J., Mihailidis, A.: Development of an automated speech recognition interface for personal emergency response systems. Journal of NeuroEngineering and Rehabilitation **6**(1) (2009) 26
9. Vacher, M., Chahuara, P., Lecouteux, B., Istrate, D., Portet, F., Joubert, T., Sehili, M.E.A., Meillon, B., Bonnefond, N., Fabre, S., Roux, C., Caffiau, S.: 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 (July 2013) 7298–7301
10. Portet, F., Vacher, M., Golanski, C., Roux, C., Meillon, B.: Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Personal and Ubiquitous Computing **17**(1) (2013) 127–144
11. López-Cózar, R., Callejas, Z.: Multimodal dialogue for ambient intelligence and smart environments. In Nakashima, H., Aghajan, H., Augusto, J.C., eds.: Handbook of Ambient Intelligence and Smart Environments. Springer US (2010) 559–579
12. Koskela, T., Väänänen-Vainio-Mattila, K.: Evolution towards smart home environments: empirical evaluation of three user interfaces. Personal and Ubiquitous Computing **8** (2004) 234–240
13. Vacher, M., Portet, F., Fleury, A., Noury, N.: Challenges in the processing of audio channels for ambient assisted living. In: IEEE HealthCom 2010, Lyon, France (Juil. 1-3 2010) 330–337
14. Mäyrä, F., Soronen, A., Vanhala, J., Mikkonen, J., Zakrzewski, M., Koskinen, I., Kuusela, K.: Probing a proactive home: Challenges in researching and designing everyday smart environments. Human Technology **2** (2006) 158–186
15. Edwards, W., Grinter, R.: At home with ubiquitous computing: Seven challenges. In Abowd, G., Brumitt, B., Shafer, S., eds.: Ubicomp 2001: Ubiquitous Computing. Volume 2201 of Lecture Notes in Computer Science., Springer Berlin / Heidelberg (2001) 256–272
16. Wölfel, M., McDonough, J.W.: Distant Speech Recognition. Wiley, New York (2009)
17. Deng, L., Acero, A., Plumpe, M., Huang, X.: Large-vocabulary speech recognition under adverse acoustic environments. In: ICSLP-2000. Volume 3., Beijing, China, ISCA (2000) 806–809
18. Baba, A., Lee, A., Saruwatari, H., Shikano, K.: Speech recognition by reverberation adapted acoustic model. In: ASJ General Meeting. (2002) 27–28
19. Michaut, F., Bellanger, M.: Filtrage adaptatif : théorie et algorithmes. Hermes Science Publication, Lavoisier (2005)
20. Valin, J.M.: On adjusting the learning rate in frequency domain echo cancellation with double talk. IEEE Transactions on Acoustics, Speech and Signal Processing **15**(3) (2007) 1030–1034
21. Vacher, M., Fleury, A., Guirand, N., Serignat, J.f., Noury, N.: Speech recognition in a smart home: some experiments for telemonitoring. In Corneliu Burileanu, H.N.T., ed.: From Speech Processing to Spoken Language Technology, Constanta (Romania), Publishing House of the Romanian Academy (2009) 171–179

22. Vacher, M., Fleury, A., Serignat, J.F., Noury, N., Glasson, H.: Preliminary Evaluation of Speech/Sound Recognition for Telemedicine Application in a Real Environment. In: Proc. InterSpeech. (2008) 496–499
23. Reidel, K., Tamblyn, R., Patel, V., Huang, A.: Pilot study of an interactive voice response system to improve medication refill compliance. BMC Medical Informatics and Decision Making **8** (2008) 46
24. Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., Bonnefond, N.: 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) 4499–4506
25. Nocéra, P., Linares, G., Massoni, D.: Phoneme lattice based A* search algorithm for speech recognition. In: LNCS : Vol. 2448, Text, Speech and Dialogue (TSD 2002), Brno, Czech Republic, Springer (2002) 301–308
26. Aman, F., Vacher, M., Rossato, S., Portet, F.: Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences. In: The 7th International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2013, Cluj-Napoca, Romania (2013) 177–184
27. Wang, Y., Zhu, X.: A new approach for incremental speaker adaptation. In: Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP 2000). (2000) 163–166
28. Fiscus, J.G.: A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In: Proc. IEEE Workshop ASRU. (1997) 347–354
29. Lecouteux, B., Linarès, G., Estève, Y., Mauclair, J.: System combination by driven decoding. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007. Volume 4. (2007) IV–341–IV–344
30. Lecouteux, B., Linarès, G., Estève, Y., Gravier, G.: Generalized driven decoding for speech recognition system combination. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008. (2008) 1549–1552
31. Lecouteux, B., Linarès, G., Nocéra, P., Bonastre, J.: Reconnaissance de la parole guidée par des transcriptions approchées. In: Journées d'Etudes sur la Parole (JEP 2006), Dinard, France (2006) 53–56
32. Berndt, D., Clifford, J.: Using dynamic time warping to find patterns in time series. In: Workshop on Knowledge Discovery in Databases (KDD'94). (1994) 359–370
33. Vacher, M., Lecouteux, B., Istrate, D., Joubert, T., Portet, F., Sehili, M., Chahuara, P.: 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) 99–105
34. Chahuara, P., Portet, F., Vacher, M.: Making context aware decision from uncertain information in a smart home: a Markov Logic Network approach. In: Ambient Intelligence. Volume 8309 of Lecture

Notes in Computer Science., Dublin, Ireland, Springer (dec 2013) 78–93
35. Franco, A.: Conférence invitée: Nouveaux paradigmes et technologies pour la santé et l'autonomie (invited conference: New paradigms and technologies for health and autonomy) [in french]. In: JEP-TALN-RECITAL 2012, Workshop ILADI 2012: Interactions Langagières pour personnes Agées Dans les habitats Intelligents (ILADI 2012: Language Interaction for Elderly in Smart Homes), Grenoble, France, ATALA/AFCP (June 2012) 1–2
36. Vacher, M., Lecouteux, B., Portet, F.: Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment. In: EUSIPCO (European Signal Processing Conference), Bucarest, Romania (August 27-31 2012) 1663–1667
37. Gemmeke, J.F., Ons, B., Tessema, N., hamme, H.V., van de Loo, J., Pauw, G.D., Daelemans, W., Huyghe, J., Derboven, J., Vuegen, L., Broeck, B.V.D., Karsmakers, P., Vanrumste, B.: Self-taught assistive vocal interfaces: an overview of the ALADIN project. In: Interspeech 2013. (2013) 2039–2043
38. Christensen, H., Casanueva, I., Cunningham, S., Green, P., Hain, T.: 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)
39. Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmueller, M., Maragos, P.: The DIRHA simulated corpus. In: The 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland (2014) 2629–2634
40. Rougui, J., Istrate, D., Souidene, W.: Audio sound event identification for distress situations and context awareness. In: Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE, Minneapolis, USA (2009) 3501–3504
41. Milhorat, P., Istrate, D., Boudy, J., Chollet, G.: Hands-free speech-sound interactions at home. In: Proceedings of the 20th European Signal Processing Conference (EUSIPCO). (aug 2012) 1678 –1682
42. Lines, L., Hone, K.S.: Multiple voices, multiple choices: Older adults' evaluation of speech output to support independent living. Gerontechnology Journal **5**(2) (2006) 78–91
43. Wolters, M.K., Georgila, K., Moore, J.D., MacPherson, S.E.: Being old doesn't mean acting old: How older users interact with spoken dialog systems. TACCESS **2**(1) (2009)
44. Cavazza, M., de la Camara, R.S., Turunen, M.: How was your day?: a companion ECA. In: AAMAS. (2010) 1629–1630
45. Istrate, D., Vacher, M., Serignat, J.F.: Embedded implementation of distress situation identification through sound analysis. The Journal on Information Technology in Healthcare **6** (2008) 204–211
46. Charalampos, D., Maglogiannis, I.: Enabling human status awareness in assistive environments based on advanced sound and motion data classification. In: Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments. (2008) 1:1–1:8

47. Popescu, M., Li, Y., Skubic, M., Rantz, M.: An acoustic fall detector system that uses sound height information to reduce the false alarm rate. In: Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008. (20–25 Aug. 2008) 4628–4631
48. Lecouteux, B., Vacher, M., Portet, F.: Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions. In Association, I.S.C., ed.: Interspeech 2011 Florence, Florence, Italy (August 2011) 2273–2276
49. Bouakaz, S., Vacher, M., Bobillier-Chaumon, M.E., Aman, F., Bekkadja, S., Portet, F., Guillou, E., Rossato, S., Desserée, E., Traineau, P., Vimon, J.P., Chevalier, T.: CIRDO: Smart companion for helping elderly to live at home for longer. IRBM **35**(2) (March 2014) 101–108
50. Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P.: The PASCAL CHiME Speech Separation and Recognition Challenge. Computer Speech and Language **27**(3) (2013) 621–633
51. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M.: The Second 'CHiME' Speech Separation and Recognition Challenge: An overview of challenge systems and outcomes. In: 2013 IEEE Automatic Speech Recognition and Understanding Workshop, Olomouc, Czech Republic (December 2013)
52. Ryan, W., Burk, K.: Perceptual and acoustic correlates in the speech of males. Journal of Communication Disorders **7** (1974) 181–192
53. Takeda, N., Thomas, G., Ludlow, C.: Aging effects on motor units in the human thyroarytenoid muscle. Laryngoscope **110** (2000) 1018–1025
54. Mueller, P., Sweeney, R., Baribeau, L.: Acoustic and morphologic study of the senescent voice. Ear, Nose, and Throat Journal **63** (1984) 71–75
55. Vipperla, R.C., Wolters, M., Georgila, K., Renals, S.: Speech Input from Older Users in Smart Environments: Challenges and Perspectives. In: HCI International: Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments. Number 5615 in Lecture Notes in Computer Science, Springer (2009) 117–126
56. Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M.S., Braga, D.: Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese. In: Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings. (2012) 139–147
57. Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K.: Acoustic Models of the Elderly for Large-Vocabulary Continuous Speech Recognition. Electronics and Communications **87**(2) (2004) 49–57
58. Vipperla, R., Renals, S., Frankel, J.: Longitudinal study of ASR performance on ageing voices. In: Proceedings of Interspeech 2008, Brisbane (2008) 2550–2553
59. Baeckman, L., Small, B. aud Whlin, A.: Handbook of the Psychology of Aging. In: Aging and memory: cognitive and biological perspectives. 5th ed. Academic Press, San Diego (2001) 349–377

60. Fozard, J., Gordont-Salant, S.: Handbook of the Psychlogy of Aging. In: Changes in vision and hearing with aging. 5th ed. Academic Press, San Diego (2001) 241–266
61. Audibert, N., Aubergé, V., Rilliard, A.: The prosodic dimensions of emotion in speech: the relative weights of parameters. In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005) 525–528
62. Vlasenko, B., Prylipko, D., Philippou-Hübner, D., Wendemuth, A.: Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In: Proceedings of Interspeech 2011. (2011) 1577–1580
63. Vlasenko, B., Prylipko, D., Wendemuth, A.: Towards robust spontaneous speech recognition with emotional speech adapted acoustic models. In: 35th German Conference on Artificial Intelligence (KI-2012), Saarbrücken, Germany (September 2012) 103–107
64. Aman, F., Auberge, V., Vacher, M.: How affects can perturbe the automatic speech recognition of domotic interactions. In: Workshop on Affective Social Speech Signals, Grenoble, France (2013) 1–5
65. Ziefle, M., Wilkowska, W.: Technology acceptability for medical assistance. In: PervasiveHealth. (March 2010) 1–9
66. McCoy, K., Waller, A.: Introduction to the special issue on AAC. ACM Transactions on Accessible Computing **1**(3) (2009)
67. McCoy, K., Arnott, J., Ferres, L., Fried-Oken, M., Roark, B.: Speech and Language processing as assistive technologies. Computer Speech and Language **27** (2013) 1143–1146