Distant Speech Recognition for Home Automation: Preliminary Experimental Results in a Smart Home

Benjamin Lecouteux, Michel Vacher, François Portet

Distant Speech Recognition for Home Automation: Preliminary Experimental Results in a Smart Home

Benjamin Lecouteux, Michel Vacher and François Portet
Laboratoire d'Informatique de Grenoble, GETALP Team
UMR CNRS/UJF/G-INP 5217
Grenoble, France
{Benjamin.Lecouteux,Michel.Vacher, Francois.Portet}@imag.fr

Abstract— **This paper presents a study that is part of the Sweet-Home project which aims at developing a new home automation system based on voice command. The study focused on two tasks: distant speech recognition and sentence spotting (e.g., recognition of domotic orders). Regarding the first task, different combinations of ASR systems, language and acoustic models were tested. Fusion of ASR outputs by consensus and with a triggered language model (using a priori knowledge) were investigated. For the sentence spotting task, an algorithm based on distance evaluation between the current ASR hypotheses and the predefine set of keyword patterns was introduced in order to retrieve the correct sentences in spite of the ASR errors.**

The techniques were assessed on real daily living data collected in a 4-room smart home that was fully equipped with standard tactile commands and with 7 wireless microphones set in the ceiling.

Thanks to Driven Decoding Algorithm techniques, a classical ASR system reached 7.9% WER against 35% WER in standard configuration and 15% with MLLR adaptation only. The best keyword pattern classification result obtained in distant speech conditions was 7.5% CER.

Keywords-component; distant speech recognition; keyword detection; triggered language models; home automation; smart home

I. INTRODUCTION

The evolution of ICT led to the emergence of smart homes equipped with ambient intelligence technology which provides high man-machine interaction capacity. Given the increase of life expectancy, these smart homes represent a promising solution to enable the elderly and frail persons to live in their own home as autonomously as possible. However, this calls for technological solutions that suit their specific needs and capabilities. Classical tactile commands may not be adapted to this population and can be complemented by speech based solutions that would provide voice command and easy interactions with their relatives or with professional carers in case of distress situations (e.g., a person who cannot move after a fall). Moreover, analysis of sounds emitted in a person's habitation may be useful for activity monitoring.

To make natural interaction with home automation possible at any time and from anywhere in the house, the Sweet-Home project was set up to integrate sound based technology within smart homes. As emphasized by Vacher *et al.* [1], major challenges still need to be overcome in these environments including robust distant speech recognition in noisy uncontrolled situations and correct identification of domotic orders in continuous audio recording conditions (i.e., the user does not press a button when she speaks to the "house"). This paper presents preliminary results of speech recognition techniques evaluated on realistic data. Before presenting how these data were acquired (Section 3), some background about the Sweet-Home project and the challenges to tackle is given in Section 2. The study consisted in two tasks. In Section 4, the first task consisting in evaluating several techniques for multi-source speech recognition is detailed. The second task, presented in Section 5, was devoted to word spotting in the recognized sentences. The paper concludes with brief remarks about the results and future work.

II. BACKGROUND

A. SWEET-HOME

The Sweet-Home project is a French national supported research project[1]. The project team is made up of researchers and engineers from the Laboratory of Informatics of Grenoble (specialized in speech processing, smart home design and evaluation), the Esigetel (specialised in audio technology) and from three companies: Theoris (real-time system development and integration), Camera Contact (diffusion and integration of adapted services for maintenance at home) and Technosens (remote health-care equipment for the elderly). The project[2] 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 her environment at any time in a most natural way.

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

[1]This work is a part of the Sweet-Home project founded by the French National Research Agency (Agence Nationale de la Recherche / ANR-09—VERS-011)

[2] http://sweet-home.imag.fr/

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 [2], a real time multi-threaded audio processing system for ubiquitous environments. The extracted information is analyzed 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). Outputs of the system thus include domotic orders, but also interaction with the user in the case when a vocal order was not understood for example, or in case of alert messages (e.g. turns off the gas, remind the person of an appointment). The system also has the possibility to make it easier for the user to connect with their relative, physician or caregiver by using the e-lio[3] or Visage[4] 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 classical 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 standardized technologies and applications. As emphasized in [3] (whose authors used X10 for their home automation bus), 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 [4]. Another example of this approach is that Sweet-Home includes systems which are already specialized 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.

B. Automatic Speech Recognition in smart homes

Automatic Speech Recognition systems (ASR) have reached good performances with close talking microphones (e.g. head-set), but the performances decrease significantly as soon as the microphone is moved away from the mouth of the speaker (e.g., when the microphone is set in the ceiling). This deterioration is due to a broad variety of effects including reverberation and presence of undetermined background noise such as TV, radio and devices. All these problems should be taken into account in the home context.

1) Echo and reverberation

Adaptation of ASR to distorted signals may be done either at the acoustic model level or at the input (feature) level [5]. Deng *et al* [6] demonstrated that adapted feature domain techniques provide better performances than those obtained by systems trained with data undergoing the same distortion as the actual data (e.g., model learned with distorted data) for both stationary and non stationary noise conditions. Moreover, for reverberation time above 500 ms, the ASR performance is not significantly improved when the acoustic models are trained on data recorded in the same reverberation condition [7]. 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 room dimensions we can assume that the reverberation time stays below 500 ms.

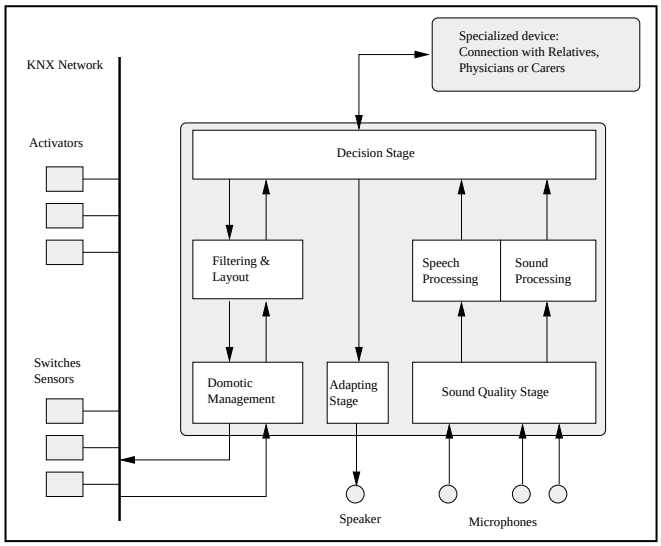

Figure 1. The SWEET-HOME diagram

2) Background noise

When the noise source perturbing the signal of interest is known, various techniques can be employed [8]. One technique is to assign a microphone to record the noise source and estimate the impulse response of the room acoustic to remove the noise [9]. This impulse response can be estimated through Least Mean Square (LMS) or Recursive Least Square (RLS) methods. In a smart home, these methods showed acceptable results when the noise is composed of speech or classical music [10]. In case of unknown noise sources, such as vacuum or ventilator, Blind Source Separation (BSS) techniques may be used. The audio signal captured by the microphone is composed of mixture of speech and noise sources.

Independent Component Analysis (ICA) is a subcategory of BSS which attempts to separate the different sources through their statistical properties (i.e., purely data driven). This method is especially efficient for non-Gaussian signals (such as speech) and does not need to take into account the position of the sources or of the microphones.

3) Sentence spotting

Term detection has been extensively studied in the last decades in the two different contexts of spoken term detection: large speech databases and keyword spotting in

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

[4] http://camera-contact.com/

continuous speech streams. The first topic recently faced a growing interest, stemming from the critical need of content-based structuring of audio-visual collections. Performances reported in the literature are quite good in clean conditions, especially with broadcast news data. Experiences undertaken in user's home conditions showed a decrease in performance. In [11], an IVR was set up to help elderly people 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 more realistic conditions, such as noisy or spontaneous speech, performances are dramatically degraded by recognition errors [12].

In this study, only homes, for which reverberation can be neglected, are considered. Therefore, only classical ASR techniques with adaptation using data recorded in the test environment are considered. For the application, some aspects of both spotting and large vocabulary continuous speech recognition (LVCSR) are encountered. Our approach is based on the use of LVCSR systems in order to increase the recognition robustness. We propose investigation in language and acoustic models adaptation and multi-source based recognition. Finally, to improve the detection rate, we propose an original approach which integrates domotic order matching directly inside the ASR system.

III. EXPERIMENTAL FRAMEWORK

To test the approach, one experiment was conducted to acquire speech corpora composed of utterances of domotic orders, distress calls and casual sentences. This corpus, called the SWEET-HOME *speech corpus*, was acquired in a real smart home using several microphones set in the ceiling. This corpus was used to tune and to test two classical ASR systems in different configurations. This section briefly introduces the corpus and the ASR systems.

A. Data acquisition in 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. To do so, the DOMUS smart home was used. 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 so that it is possible to act on the sensory ambiance, depending on the context and the user's habits. The flat is fully usable and can accommodate a dweller for several days. The technical architecture of DOMUS is based on the KNX bus system (KoNneX)[5], a worldwide ISO standard (ISO/IEC 14543) for home and building control. More than 150 sensors, actuators and information providers are managed in the flat. For the need of the SWEET-HOME project, the flat has also been equipped with 7 radio microphones set into the ceiling (2 per room except for the bathroom) that can be recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card [1]. A full record of the sound flow of the 7 channels was done for each speaker during the entire experiment.

Figure 2. The Domus Smart Home

From October 1 to November 30 2010, 21 persons (including 7 women) participated to a 2-phase experiment to record, among other data, speech corpus in a daily living context. To make sure that the audio data acquired would be as close as possible to real daily living sounds, the participants were asked to perform several daily living activities in the smart home. The average age of the participants was 38.5±13 years (22-63, min-max) and each experimental session lasted about 2 hours. No instruction was given to any participant about how they should speak and in which direction. Consequently, no participant emitted sentences directing their voice to a particular microphone.

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, simulate a shower, get some sleep, clean up the flat using the vacuum, etc.). A visit, before the experiment, was organized to make sure that the participants will find all the items necessary to perform the activities. During this first phase, participants uttered forty predefined casual sentences on the phone (e.g., "Allo", "J'ai eu du mal à dormir") but were also free to utter any sentence they wanted (some did speak to themselves aloud). Note that only audio information from the seven microphones on the ceiling was captured and not from the microphone of the telephone. Data from video cameras and the domotic network were also captured but their descriptions are out of the scope of this paper.

The second phase consisted in reading aloud a list of 44 sentences whose 9 were distress sentences and 3 were domotic orders. This list was read in 3 rooms (study, bedroom, and kitchen) under three conditions: with the vacuum on, with the radio on (vacuum off) and without noise (everything off). So each participant read the text nine times in total, which makes 396 sentences. Only the clean condition will be used in this paper, the noisy condition records having been designed for other experiments.

B. SWEET-HOME speech corpus

For the phase 1, only the sentences uttered in the study during the phone conversation were considered. For the

[5] www.knx.org

phase 2, only the sentences uttered in the kitchen without additional noise (vacuum or radio) were considered. This corpus was indexed manually because each speaker did not follow strictly the instructions given at the beginning of the experiment. Some hesitations and word repetitions occurred along the records. Moreover, when two sentences were uttered without a sufficient silence between them, some of these couples were considered as one sentence. A summary of the corpus is given in Table 1. The SWEET-HOME speech corpus is made of 862 French sentences uttered by 21 persons in the first phase, 917 French 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 average SNR (Signal-to-Noise Ratio) for the considered sentences is 20.3 dB.

C. ASR systems

Two classical ASR systems were used in the study: Sphinx [13] and Speeral [14]. This section introduces both system and the French language models used. We used previously computed French acoustic models.

1) The Speeral ASR system

The LIA (Laboratoire d'Informatique d'Avignon) speech recognition tool-kit Speeral relies on a decoder with HMM-based context-dependent acoustic models and trigram language models. HMMs are classical three-state left-right models and state tying is achieved by using decision trees. Acoustic vectors are composed of 12 PLP (Perceptual Linear Predictive) coefficients, the energy, and the first and second order derivatives of these 13 parameters. Speeral was used to build a system dedicated to broadcast news; it was involved in the ESTER evaluation campaign [14]. In the study, the acoustic models were trained on ESTER materials (about 80 hours of annotated speech). Given the targeted application of SWEET-HOME the computation time should not be a breach of real-time use. Thus, the 1xRT Speeral configuration was used. In this case, the time required by the system to decode one hour of speech signal is real-time (noted 1xRT). The 1xRT system uses a strict pruning scheme.

2) The Sphinx ASR system

The CMU Sphinx 3.3 (fast) decoder [13] is a branch of the CMU Sphinx III project which has been developed to increase the speed of the algorithm. This decoder uses fully continuous acoustic models with 3 or 5-state left-to-right HMM topologies. Acoustic vectors are composed of 13 MFCC coefficients, the delta and the double delta of each coefficient. HMM-based context-dependent acoustic models were trained on the BREF 120 corpus [15] which is composed of about 100 hours of annotated speech from 120 French speakers. We also attempted to use ESTER data for the training; however results were much degraded and are not presented in this paper.

3) Language models

For each ASR, the acoustic model was different but the same 3-gram language model with a 10K lexicon was used with both. Two language models were considered in the study: the generic and the specialized models. The *generic* language model 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* language model was estimated from the sentences that the 21 participants had to utter during the experiment (domotic orders, casual phrases, etc.).

TABLE I. SWEET-HOME SPEECH CORPUS DESCRIPTION

Spkr. ID	Phase 1			Phase 2
	Duration (s) Channel 6 or 7	*SNR mean (dB) Channel 6*	*SNR mean (dB) Canal 7*	*Duration (s) Channel 4 or 5*
1	145.78	23.54	22.08	96.66
2	119.36	22.64	21.04	110.42
3	112.08	14.80	12.21	119.76
4	141.32	16.51	16.47	119.04
5	159.32	29.70	26.75	122.21
6	122.10	17.65	16.11	108.61
7	110.90	19.04	17.52	116.00
8	114.54	20.31	18.99	114.64
9	121.58	26.80	24.73	135.36
10	77.50	20.27	18.00	104.54
11	106.52	20.22	21.06	105.76
12	90.48	24.50	21.08	108.44
13	96.46	26.24	19.88	116.52
14	97.74	17.69	17.66	113.40
15	96.48	22.55	21.36	101.98
16	96.86	21.39	17.61	106.72
17	111.08	21.66	20.00	144.46
18	169.14	19.97	19.04	124.52
19	146.98	25.12	23.41	125.58
20	89.80	27.46	24.77	120.60
21	99.48	19.46	19.18	109.56

IV. SPEECH RECOGNITION TASK

The two ASR systems (Sphinx and Speeral) performances were compared in 1xRT conditions. In order to propose **a baseline system**, the adaptation of both acoustic and language models was tested respectively on Sphinx and Speeral. Then, to improve the robustness of the recognition, multi-source ASR was tested. Finally, a new variant of a driven decoding algorithm was used in order to take into account available *a priori* information and several audio channels for each speaker.

In each case, the phase 1 of the corpus was used for development and MLLR speaker adaptation while the phase 2 was used for performances estimation. Results obtained on the phase 2 of the corpus were compared at two levels: the Word Error Rate (WER) and the Classification Error Rate (CER). The WER is a good measure for the robustness, while the CER corresponds to the main goal of our research (i.e., detection of *predefined* sentences).

A. Acoustic models adaptation: MAP versus MLLR

To improve the recognition, acoustic models were adapted for each speaker by using two methods: Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR). This was achieved by using data of the

first phase. These data were perfectly annotated, allowing performing 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. The principle is that 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.

The figure 3 shows results for the ASR systems without adapted acoustic models. All experiments are carried out with the generic language model (GLM) lightly interpolated with predefined sentences (PS) presented in the next section (GLM: 90%, PS: 10%). Without acoustic adaptation, the mean WER is 57.7% for Sphinx and 35% for Speeral.

The figure 4 presents the results using MAP adaptation. They show that MAP is not very relevant in this particular case. With MAP, the Speeral WER is 28.5% and the Sphinx WER is 62%. Two aspects explain this:

- The lack of parameter tying in the standard MAP algorithm implies that the adaptation is not robust.
- The noisy environment is not suited to MAP adaptation [21].

The best results are obtained with MLLR adaptation, which is the best choice for sparse and noisy adaptation data whatever the channel. The results are presented in figure 5. The mean Speeral WER is 14.5% while the Sphinx WER is 27.3%.

B. Reducing linguistic variability

In this section, we propose some linear interpolation schemes 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. Better recognition can be obtained by reducing the overall linguistic space. This can be achieved by estimating a language model on the expected domotic orders. However, such a language model would be probably too specific when the speaker deviates from the original transcript.

Therefore, we interpolated a specialized model with a generic large vocabulary language model.

Two schemes of linear interpolation were considered: in the first one, the *generic* model had a strong weight; in the second one, the impact of the *generic* model was low. The results of ASR in four conditions — generic language model only, specialized model only, and the two interpolations — are presented below. The ASRs were assessed after MLLR adaptation using the data of phase 1 of the corpus.

The figure 6 presents ASR systems WER with the generic language model (Baseline). As expected, the baseline language model obtained poor results: about 85.2% of WER for Sphinx and 75.3% for Speeral. Without reliable information, the ASR systems, in noisy, speaker independent and large vocabulary condition were unable to perform good recognition.

The figure 7 presents ASR systems WER with the specialized language model. The systems were 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 language model is thus too specific.

Finally, a light (10%) interpolated language model led to the best results, which are presented in Figure 8. This model combined the generic language model (with a 10% weight) and the specialized 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. However, the best WER is obtained when a generic language model is also considered: when the speaker deviates, the generic language model allows to correctly recognize the pronounced sentences.

C. Conclusion about monosource ASR with Sphinx and Speeral

Sphinx and Speeral were assessed taking into account realistic distant-speech conditions and a home automation application (voice command). Thus, each of the systems had to perform ASR with several constraints and opportunities. Indeed, the noisy, distant-speech, multispeaker (more than one person may be in the home), continuous analysis and real-time aspects put the experiment in more difficult conditions than the classical head-set one. In the experiment, the two ASR systems did not give the same performances. In this special case, we observe that Speeral led to the best results. However these differences must be further explored, because acoustic models were not trained on the same data set.

In our next experiments, the selected baseline system is based on Speeral with a specific MLLR adaptation for each speaker and an interpolated language model.

The application conditions also make the ASR systems benefit from multiple audio channels, from a reduced vocabulary and from the hypothesis that only one speaker should utter voice commands. On the two systems, lightly interpolated language model and a MLLR acoustic adaptation did improve significantly the ASR system performance, with Speeral still leading to the lowest WER. In the next section, we propose several techniques to perform multi-source ASR system and to combine systems.

D. ROVER

The ROVER algorithm [16] (Recognizer Output Voting Error Reduction) allows voting methods to be used for word-level system combination within large vocabulary speech recognition tasks. This scheme makes use of the one-best hypothesis from a set of speech recognizers, with optional confidence associated for each output. The ROVER vote is an easy way to combine multiple systems. In this study, ROVER was used to benefit from a potential

complementarity of the ASR systems. In our case, confidence scores were not associated to system outputs: a majority vote was used.

Results of the ROVER applied to the experimental data in multiple ASR systems configurations are presented in figure 9. We combined all systems:

- Speeral MLLR streams 1 & 2
- Sphinx MLLR streams 1 & 2

Stream 1 is related to channel 4 and stream 2 to channel 5.

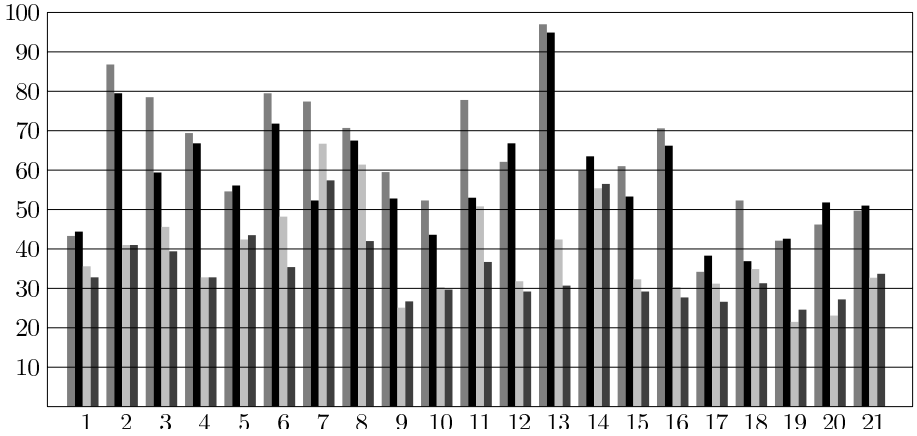

Figure 3. ASR system recognition WER without acoustic adaptation. From left to right: two Sphinx streams and two Speeral streams

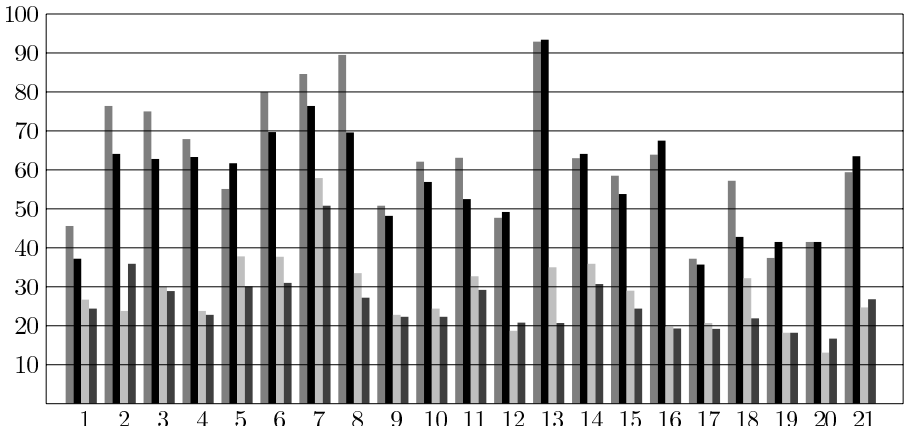

Figure 4. ASR system recognition WER with MAP acoustic adaptation. From left to right: two Sphinx streams and two Speeral streams

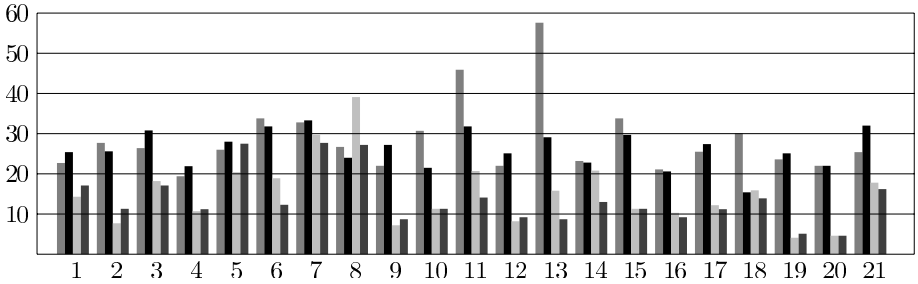

Figure 5. ASR system recognition WER with MLLR acoustic adaptation. From left to right: two Sphinx streams and two Speeral streams

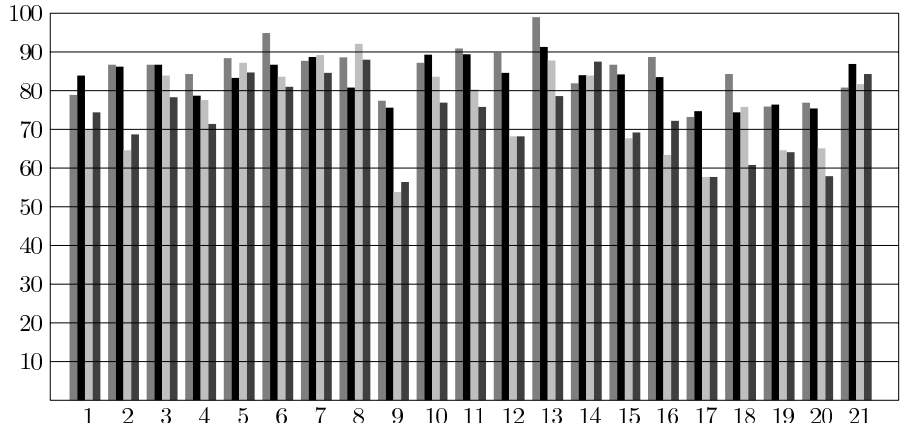

Figure 6. ASR system recognition WER with the baseline Language Model. From left to right: two Sphinx streams and two Speeral streams

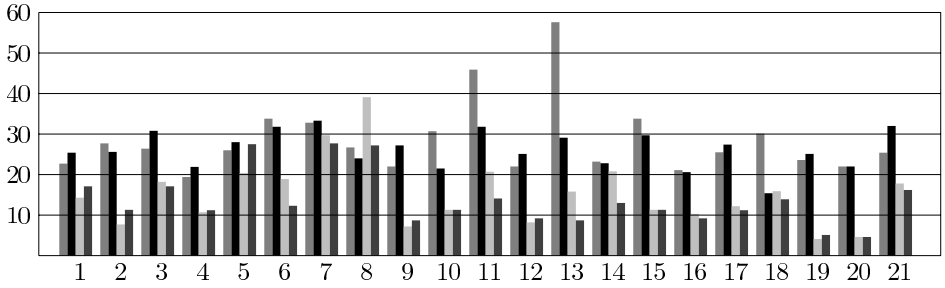

Figure 7. ASR system recognition WER with LM trained on a priori sentences keywords. From left to right: two Sphinx streams and two Speeral streams

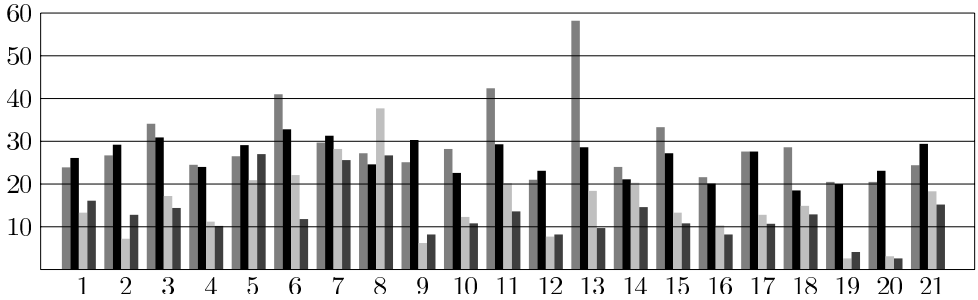

Figure 8. ASR system recognition WER with the specilized LM. From left to right: two Sphinx streams and two Speeral streams

Despite of the poor Sphinx performance, the ROVER combination led to great improvements. The results show that the ROVER makes possible robust ASR with a mean WER of 10.2%: this aspect show the complementarity of the systems and the streams. However, the ROVER stage increased the computation time by the number of combined systems. 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 systems reduces the problem to picking the word with the highest confidence when two systems disagree. Thus, when the recognizer confidence scores are not reliable, the ROVER between two systems 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. Then ROVER results are used as baseline in next experiments.

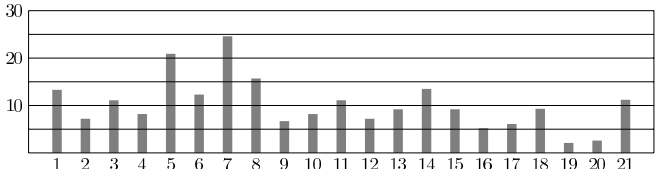

Figure 9. ASR system recognition WER by using ROVER

E. Driven Decoding Algorithm

We recently proposed the Driven Decoding Algorithm (DDA) [17, 18] which is able to simultaneously align and correct imperfect ASR outputs [19]. 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 with the language model:

$$\widetilde{P}\left(w_i|w_{i-1},w_{i-2}\right)=P^{1-\alpha}\left(w_i|w_{i-1},w_{i-2}\right) \qquad (1)$$

where $\widetilde{P}\left(w_i|w_{i-1},w_{i-2}\right)$ is the updated trigram probability of the word given the history w_{i-2}, w_{i-3} and $P\left(w_i|w_{i-1},w_{i-2}\right)$ 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 10. 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 0.1xRT)
- While a ROVER does not allow combining efficiently two systems without confidence scores, DDA combines easily the information.

The Figure 10 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 combining the information of the two streams.

Results using the two stream-DDA are presented in Figure 11. In most cases, DDA generated hypothesis that led either to the mean WER of the two initial streams or to better WER. These results underline an interesting feature of the DDA in comparison with a simple ROVER to combine two systems: the mean WER is 12.3%. We propose to extend this approach in the next section by driving the ASR system by *a priori* sentences selected on the first stream.

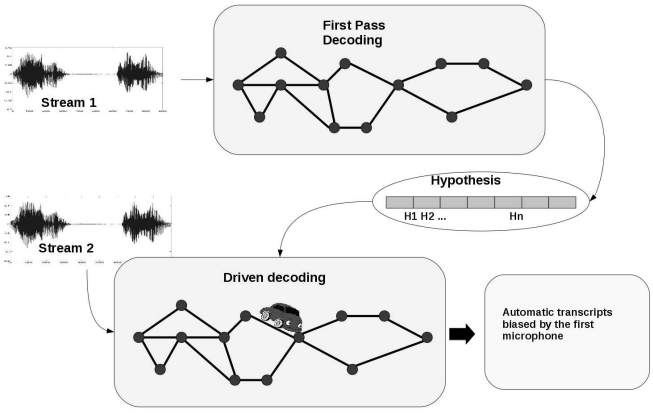

Figure 10. Driven Decoding Algorithm used with two streams: The first stream allows one to drive the second stream

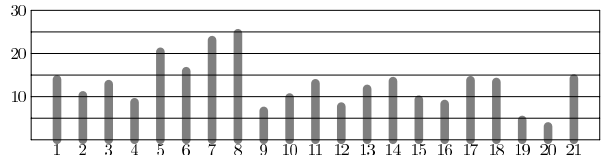

Figure 11. ASR system recognition WER by using DDA

F. *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 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 [22]: 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 generating hypotheses. By using the edit distance explained in [22], closed spotted sentences are selected and used as input for the fast second pass as presented in Figure 12. In this configuration, the first pass is used to select some sentences used to drive the second pass. In the Figure 12, the first system output is " allumer la lumière". The edit distance allowed finding two closed sentences: "allumez la lumière" and "allumez la télévision". These sentences drive the second pass and allow one to find the good output "allumez la lumière".

Results using this approach are showed in Figure 13. 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 7.9%.

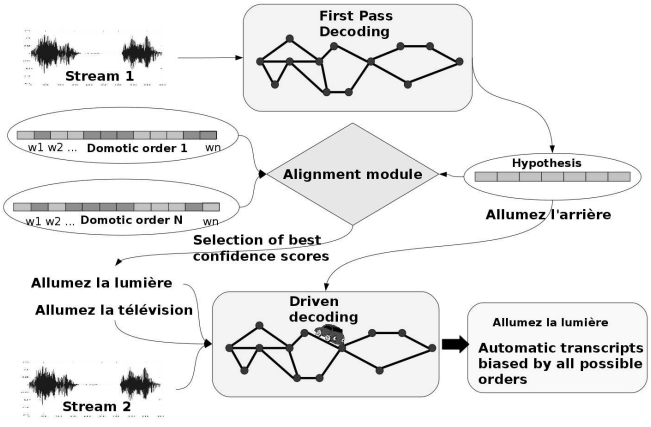

Figure 12. Driven Decoding Algorithm used with two streams and a priori sentences. The first stream allows one to drive the second stream, according to a refine selection of spotted sentences.

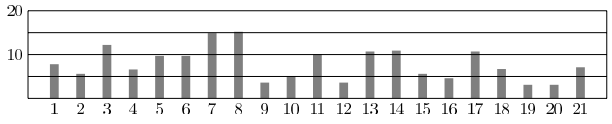

Figure 13. ASR system recognition WER by using the two-level DDA

G. Two level DDA: conclusion

In the case of constrained computational resources, the ASR system Speeral seems more accurate than the ASR system Sphinx. Nevertheless, the ROVER between all systems shows that a part of information is lost. Then, we proposed two original approaches based on the Driven Decoding Algorithm. 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.

V. DETECTION OF PREDEFINED SENTENCES

In order to spot sentences into automatic transcripts T of m characters, each sentence of n characters from *predefined* sentences H was aligned with T by using a Dynamic Time Warping (DTW) algorithm at the letter level [20]. Sequences were aligned by constructing an n-by-m matrix where the element of the matrix contained the distance between the two words and using the distance function defined below.

$$\begin{aligned} &d(T_i,H_j)=0 \text{ if } T_i=H_j \\ &d(T_i,H_j)=3 \text{ in the insertion cases} \\ &d(T_i,H_j)=3 \text{ in the deletion cases} \\ &d(T_i,H_j)=6 \text{ in the substitution cases} \end{aligned} \quad (2)$$

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

$$\gamma(i,j)=d(T_i,H_j)+\min\left(\gamma(i-1,j-1),\gamma(i-1,j),\gamma(i,j-1)\right) \quad (3)$$

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 declinations or light variations ("téléviseur", "télévision", etc.). Moreover, in a lot of cases, a miss-decoded word is 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" "*turn on the light*"), the detection methods were applied in the following ASR configurations:

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

The Figure 14 presents the correct detected predefined sentences (e.g., $1-CER$) in the 4 configurations. As expected, the three systems based on ROVER and DDA gave the best performances, with respectively 88.2%, 87,4% and 92.5% of correct classifications. 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.

A. Spotting task: conclusion

In all best-configurations, *predefined* sentence recognition had a good accuracy: the baseline recognition gave 83.1%. It can be observed that in other configurations the spotting task correlated 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.

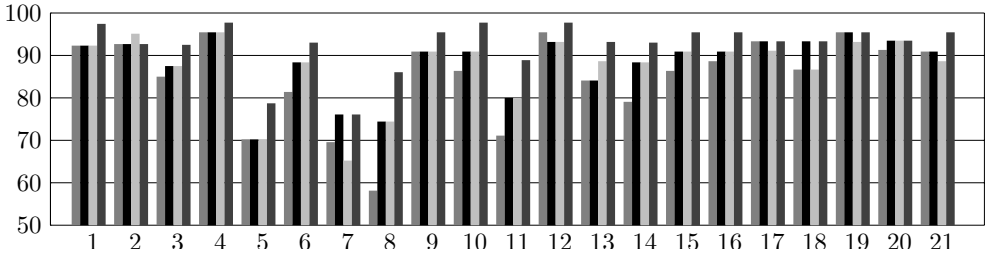

Figure 14. ASR system recognition in term of sentence spotting. From left to right: baseline, ROVER, DDA1, DDA2

VI. CONCLUSION

This paper describes a study of speech recognition and sentence spotting in a smart home implying noisy and distant speech conditions. Two ASR systems: Sphinx and Speeral working with different graph exploration algorithms were used to build the baseline system. The best baseline system was composed of interpolated language models between a large vocabulary one and a specialized one (composed of predefined domotic sentences); and used MLLR acoustic adaptation. The best baseline obtained a 14.5% WER and a 16.9% CER.

In a second part, an original approach was proposed to benefit from the multiple microphones of the smart home and *a priori* knowledge about the sentences being uttered. This approach is base on the Driven Decoding Algorithm which permits to drive a stream being decoded by the results of the decoding on another one. This approach was refined by integrating *a priori* knowledge in the DDA about the sentences to spot. Experimental results showed that these approaches brought significant gain compared to the baseline: the best results were obtained by a two-level DDA. DDA led to both a WER (7.9%) and CER (7.5%) improvement. Moreover, results were better than the ones of a ROVER based system working with more than 5 times real time while the two-level DDA approach worked in about 1xRT. In fact, DDA benefits from the *a priori* knowledge to speed up the speech recognition by reducing the search space.

This study shows that good recognition rate can be obtained by adapting classical ASR systems mixing multi-source and domain knowledge. However, in a smart home many unknown audio sources can perturb the speaker voice. This noise must be removed to permit the system being efficient even in noisy condition. We plan to test these approaches in noisy conditions with Independent Component Analysis (ICA) in order to separate speech from undetermined noise.

ACKNOWLEDGMENT

The authors would like to thank S. Méniard who developed the StreamHIS software used for multichannel sound recording. They also are very grateful to the participants who took part to the different experiments. Thanks are also extended to B. Meillon, N. Bonnefond and S. Pons for their support during the experiment.

REFERENCES

[1] 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(1), pp. 35-54, January-March 2011.

[2] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, and N. Noury, *Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living.* Intech Book, 2010, pp. 645 – 673.

[3] F. Mäyrä, A. Soronen, J. Vanhala, J. Mikkonen, M. Zakrzewski, I. Koskinen, and K. Kuusela, "Probing a proactive home: Challenges in researching and designing everyday smart environments," *Human Technology*, vol. 2, pp. 158–186, 2006.

[4] W. Edwards and R. Grinter, "At home with ubiquitous computing: Seven challenges," in *Ubicomp 2001: Ubiquitous Computing*, ser. Lecture Notes in Computer Science, vol. 2201. Springer Berlin / Heidelberg, 2001, pp. 256–272.

[5] M. Wölfel and J. W. McDonough, *Distant Speech Recognition*. New York: Wiley, 2009.

[6] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *ICSLP-2000*, vol. 3. Beijing, China: ISCA, 2000, pp. 806–809.

[7] A. Baba, A. Lee, H. Saruwatari, and K. Shikano, "Speech recognition by reverberation adapted acoustic model," in *ASJ General Meeting*, 2002, pp. 27–28.

[8] F. Michaut and M. Bellanger, *Filtrage adaptatif : théorie et algorithmes*. Hermes Science Publication, Lavoisier, 2005.

[9] J.-M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double talk," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 15, no. 3, pp. 1030–1034.

[10] M. Vacher, A. Fleury, N. Guirand, J.-f. Serignat, and N. Noury, "Speech recognition in a smart home: some experiments for telemonitoring," in Proc. *SPED, From Speech Processing to Spoken Language Technology*, Constanta (Romania), 2009, pp. 171–179.

[11] K. Reidel, R. Tamblyn, V. Patel, and A. Huang, "Pilot study of an interactive voice response system to improve medication refill compliance," *BMC Medical Informatics and Decision Making*, vol. 8, p. 46, 2008.

[12] M. Vacher, A. Fleury, J.-F. Serignat, N. Noury, and H. Glasson, "Preliminary Evaluation of Speech/Sound Recognition for Telemedicine Application in a Real Environment," in Proc. *9th International Conf. on Speech Science and Speech Technology* (InterSpeech 2008), vol. 1, Brisbane (Australia), 2008, pp. 496–499.

[13] K. Seymore, C. Stanley, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer, "The 1997 cmu sphinx-3 english broadcast news transcription system," in *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.

[14] P. Nocera, G. Linares, and D. Massonié, "Principes et performances du décodeur parole continue speeral," in *XXIVées journées d'étude sur la parole*. Laboratoire Informatique d'Avignon, 2002.

[15] J.-L. Gauvain, L.-F. Lamel, and M. Eskenazi, "Design considerations and text selection for BREF, a large french read-speech corpus," in Proc. *International Conference on Spoken Language Processing*, Kobe, Japan, 1990, pp. 1097–1100.

[16] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 347–354.

[17] B. Lecouteux, G. Linarès, F. Beaugendre, and P. Nocéra, "Text island spotting in large speech databases," in Proc. *Interspeech'07*, 2007, pp. 1318–1321.

[18] B. Lecouteux, G. Linarès, Y. Estève, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing* ICASSP 2008, 2008, pp. 1549–1552.

[19] B. Lecouteux, G. Linarès, J. Bonastre, and P. Nocéra, "Imperfect transcript driven speech recognition," in Proc. *InterSpeech'06*, Pittsburg (USA), 2006, pp. 1626–1629.

[20] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Workshop on Knowledge Discovery in Databases (KDD'94)*, 1994, pp. 359–370.

[21] Y. Wang and X. Zhu, "A new approach for incremental speaker adaptation", in Proc. *International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 163-166.

[22] B. Lecouteux, M. Vacher and F. Portet , "Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions", in Proc. *Interspeech 2011*, Florence, Italy, Aug. 2011, 4p.