Sound Environment Analysis in Smart Home

Mohamed A. Sehili[1,3], Benjamin Lecouteux[2], Michel Vacher[2], François Portet[2], Dan Istrate[1], Bernadette Dorizzi[3], and Jérôme Boudy[3]

[1] ESIGETEL, 1 Rue du Port de Valvins, 77210 Avon - France
{mohamed.sehili,dan.istrate}@esigetel.fr
[2] Laboratoire d'Informatique de Grenoble
Grenoble 1/Grenoble INP/CNRS UMR 5217, F-38041 Grenoble - France
{benjamin.lecouteux,francois.portet,michel.vacher}@imag.fr
[3] Telecom SudParis, 9 Rue Charles Fourier, 91000 Évry - France
{bernadette.dorizzi,jerome.boudy}@it-sudparis.eu

Résumé This study aims at providing audio-based interaction technology that lets the users have full control over their home environment, at detecting distress situations and at easing the social inclusion of the elderly and frail population. The paper presents the sound and speech analysis system evaluated thanks to a corpus of data acquired in a real smart home environment. The 4 steps of analysis are signal detection, speech/sound discrimination, sound classification and speech recognition. The results are presented for each step and globally. The very first experiments show promising results be it for the modules evaluated independently or for the whole system.

Keywords: Smart Home, Sound Analysis, Sound Detection, Sound Recognition, Speech Distant Recognition

1 Introduction

1.1 General aspects

Demographic change and aging in developed countries imply challenges for the society to continue to improve the well being of its elderly and frail inhabitants. Since the dramatic evolution of Information and Communication Technologies (ICT), one way to achieve this aim is to promote the development of smart homes. In the health domain, a *health smart home* is a habitation equipped with a set of sensors, actuators, automated devices to provide ambient intelligence for daily living task support, early detection of distress situations, remote monitoring and promotion of safety and well-being [1]. The smart home domain is greatly influenced by the *Ubiquitous Computing* domain. As introduced by Weiser [2], ubiquitous computing refers to the computing technology which disappears into the background, which becomes so seamlessly integrated into our environment that we do use it naturally without noticing it. Among all the interaction and sensing technologies used in smart homes (e.g., infra-red sensors, contact doors, video cameras, RFID tags, etc.), audio processing technology has a great potential to become one of the major interaction modalities.

Audio technology has not only reached a stage of maturity but has also many properties that fit the Weiser's vision. It is physically intangible and depending on the number and type of the sensors (omnidirectional microphones) that are used, it does not force the user to be physically at a particular place in order to operate. Moreover, it can provide interaction using natural language so that the user does not have to learn complex computing procedures or jargon. It can also capture sounds of everyday life which makes it even more easy to use and can be used to communicate with the user using synthetic or pre-recorded voice. More generally, voice interfaces can be much more adapted to disabled people and the aging population who have difficulties in moving or seeing than tactile interfaces (e.g., remote control) which require physical and visual interaction. Moreover, audio processing is particularly suited to distress situations. A person, who cannot move after a fall but being conscious, still has the possibility to call for assistance while a remote control may be unreachable. Despite all this, audio technology is still underexploited. Part of this can be attributed to the complexity of setting up this technology in a real environment and to important challenges that still need to be overcome [3].

1.2 Related work

Compared to other modalities (e.g., video cameras, RFID tags), audio technology has received little attention [4,5,6,7,8,9,10]. To the best of our knowledge, the main trends in audio technology in smart home are related to dialog systems (notably for virtual assistant/robot) and emergency and the main applications are augmented human machine interaction (e.g., voice command, conversation) and security (mainly fall detection and distress situation recognition).

Regarding security in the home, audio technology can play a major role in smart homes by helping a person in danger to call for help from anywhere without having to use a touch interface that can be out of reach [4,10]. Another application is the fall detection using a the signal of a wearable microphone which is often fused with other modalities (e.g., accelerometer) [6,5]. In [5] a patient awareness system is proposed to detect body fall and vocal stress in speech expression through the analysis of acoustic and motion data (microphones and accelerometers). However, the person is constrained to wear these sensors all the time. To address this constraint, the dialog system developed by [8] was proposed to replace traditional emergency systems that requires too much change in lifestyle of the elders. However, the prototype had a limited vocabulary (yes/no dialog), was not tested with aged users and there is no mention about how the noise was taken into account. In [9], a communicative avatar was designed to interact with a person in a smart office. In this research, enhanced speech recognition is performed using beam-forming and a geometric area of recording. But this promising research is still to be tested in a multiroom and multisource realistic home.

Most of the speech related research or industrial projects in AAL are actually highly focused on dialog to build communicative agent (e.g., see the EU funded

Companions or CompanionAble projects or the Semvox system [4]). These systems are often composed of Automatic Speech Recognition, Natural Language Understanding, Dialog management, and Speech Synthesis parts supplying the user the ability to communicate with the system in an interactive fashion. However, it is generally the dialog module (management, modeling, architecture, personalization, etc.) that is the main focus of these projects (e.g., see Companions, OwlSpeak or Jaspis). Moreover, this setting is different from the SWEET-HOME one as the user must be close to the avatar to speak (i.e., not a distant speech setting). Indeed, in SWEET-HOME, the aim is to enable speech interaction from anywhere in multiroom-home. Furthermore, few research projects have considered using daily living sound in their systems though it can be useful information [11,4]. In this perspective, the project addresses the important issues of distant-speech and sound source identification and the outcomes of this research is of high importance to improve the robustness of the systems mentioned above.

1.3 The SWEET-HOME Project

SWEET-HOME is a project aiming at designing a new smart home system based on audio technology 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, from anywhere in the house, their environment at any time in the most natural possible way. The targeted users are elderly people who are frail but still autonomous. There are two reasons for this choice. Firstly, a home automation system is costly and would be much more profitable if it is used in a life-long way rather than only when the need for assistance appears. Secondly, in the case of a loss of autonomy, the person would continue to use their own system with some adaptations needed by the new situation (e.g., wheelchair) rather than having to cope simultaneously with their loss of autonomy and a new way of life imposed by the smart home. Qualitative user evaluation studies showed that such speech technology is well accepted by the elderly people [12].

2 Proposed sound analysis system

The proposed sound processing system (Figure 1) is composed of the following 4 stages which will be described in the next subsections :

1. the **Sound Detection and Extraction** stage, detecting sound events (daily sounds or speech) from input streams ;
2. the **Sound/Speech Discrimination** stage, discriminating speech from other sounds to extract voice commands ;
3. the **Sound Classification** stage, recognizing daily living sounds ; and
4. the **Speech Recognition** stage, applying speech recognition to events classified as speech.

[4] `http://www.semvox.de`

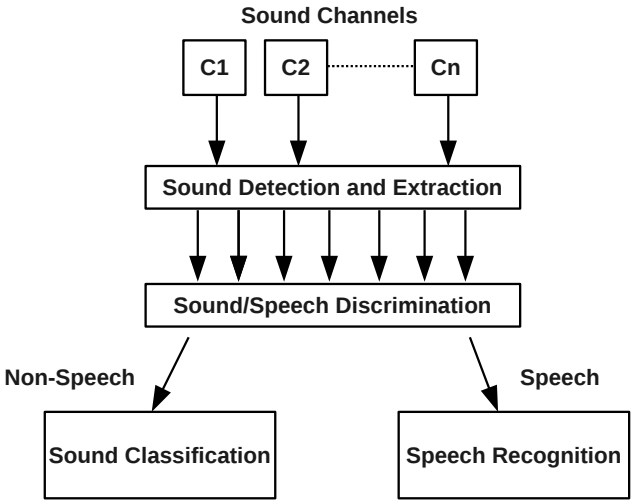

Fig. 1. Sound Analysis System in the SWEET-HOME Project.

2.1 Sound Detection and Extraction

To detect intervals of sound occurrence an algorithm based on Discrete Wavelet Transform (DWT) is applied. The algorithm computes the energy of the three high frequencies coefficients of the DWT and an auto-adaptive threshold is estimated from the current Signal to Noise Ratio (SNR). In this work the SNR is computed based on the hypothesis that the noise present within the sound event interval is similar to the noise in the frames directly preceding the current event. The estimation of the SNR is useful both for adapting the threshold and for rejecting events too noisy to be treated. For more details, the reader is referred to [13].

2.2 Sound/Speech Discrimination and Sound Classification

Once intervals of sound occurrences are detected, the most important task is to recognize their category. In everyday life, there is a large number of different sounds and modeling all of them is intractable and not necessary in our case. Indeed, given that the primary aim for the SWEET-HOME project is to enable distant voice command and distress situation detection, speech is the most important sound class. However, other kinds of sound are of interest such as crying and screaming which might reveal that the person needs urgent help. Moreover, some environmental sounds such as glass breaking and water flowing for a long while could also indicate a situation that requires external intervention.

These facts motivated us to build a hierarchical sound recognition system. First, speech is differentiated from all other sounds using a speech-against-sound

model rather than including speech as one of the classes of interest. The two-class classification strategy is generally more reliable than numerous-class classification schemes. Then a multi-class sound classification is performed in order to recognize non-speech sounds.

We use the same method for sound/speech discrimination and sound classification. It is a combination of two well-known methods, GMM (Gaussian Mixture Models) and SVM (Support Vector Machines), and it belongs to the so called sequence discriminant kernels [14] [15]. Sequence discriminant kernels for SVM were successfully used for speaker recognition and verification and they became a standard tool [16]. Their main advantage is their ability to deal with sequences of observations of variable length. A sequence of vectors is classified as a whole without performing frame-level classification as with GMM. This makes them quite suitable for sound classification as sound duration may greatly vary. The kernel used in this work is called the GMM Supervector Linear Kernel (GSL) [17] [18]. The following subsections explain this kernel in details.

Support Vector Machines Support Vector Machines [19] [20] is a discriminative method often used for classification problems, especially when dealing with non-linear data spaces. The non-linear classification is achieved thanks to kernel functions :

$$f(\mathbf{x}) = \sum_{i=1}^{N_{sv}} \alpha_i y_i K(\mathbf{x}, \mathbf{x_i}) + b \tag{1}$$

where $\mathbf{x_i}$ are the support vectors chosen from training data via an optimization process and $y_i \in \{-1, +1\}$ are respectively their associated labels. $K(.,.)$ is the kernel function and must fulfill some conditions so that : $K(\mathbf{x}, \mathbf{y}) = \mathbf{\Phi}(\mathbf{x})^t \mathbf{\Phi}(\mathbf{y})$ where $\mathbf{\Phi}$ is a mapping from the input space to a possible infinite-dimensional space.

Using SVM at frame-level for sound classification is very time consuming both for training and recognition processes and leads to low performances [21]. Thus, sequence discrimination kernels are used to overcome these problems. In [18] the following general definition of a sequence kernel is proposed :

$$K(X, Y) = \mathbf{\Phi}(X)^t \mathbf{R^{-1}} \mathbf{\Phi}(Y) = \left(\mathbf{R}^{-\frac{1}{2}} \mathbf{m}^X\right)^t \left(\mathbf{R}^{-\frac{1}{2}} \mathbf{m}^Y\right) \tag{2}$$

where $\mathbf{\Phi}(X)$ is the high-dimensional vector resulting from the mapping of sequence X and $\mathbf{R^{-1}}$ is a diagonal normalization matrix. In [18] a comparison between two sequence kernels for speaker verification was made. The GSL kernel showed better performances than the Generalized Linear Discriminant Sequence Kernel (GLDS) and thus we retained it for our sound classification system.

GMM Supervector Linear Kernel The GSL scheme is depicted in Figure 2. To compute the kernel K as in equation (2), we define as $\mathbf{\Phi}_{\mathrm{GSL}}(X) = \mathbf{m}^X$ the supervector composed of the stacked means from the respective adapted Universal Background Model (UBM) components. For each sequence of vectors

X extracted from one sound event, a GMM UBM of diagonal covariance matrices is adapted via a MAP (Maximum *a posteriori*) procedure [22]. The normalization matrix $\mathbf{R^{-1}}$ is defined using the weights and the covariances of the UBM model. For more details about the GSL procedure, the reader is referred to [18].

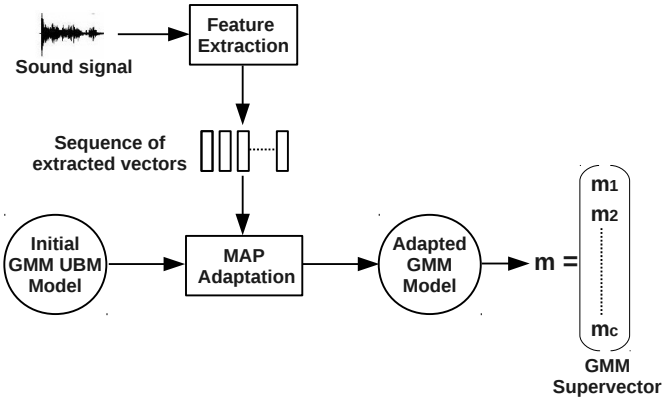

Fig. 2. GMM Supervector mapping process

2.3 Speech Recognition

Automatic Speech Recognition systems (ASR) have reached good performances with close talking microphones (e.g. head-set), but the performance decreases 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 and have become hot topics in the speech processing community [23].

In the Sweet-Home project, only vocal orders, large speech or some distress sentences need to be detected. 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 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. However, performances of state-of-the-art approach are unknown in noisy situation such as the one considered in Sweet-Home. This section summarizes experiments that were run to test to which extend standard and research ASR systems can be used in this context.

The LIA (Laboratoire d'Informatique d'Avignon) speech recognition toolkit Speeral [24] was chosen as unique ASR system. Speeral relies on an A$\star$

decoder with Hiden Markov Models (HMM) based context-dependent acoustic models and trigram language models. HMMs use 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. In the study, the acoustic models were trained on 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. Furthermore, acoustic models were adapted for each of the 21 speakers by using the Maximum Likelihood Linear Regression (MLLR) and the annotated Training Phase of the corpus. MLLR adaptation is a good compromise while only a small amount of annotated data is available. For the decoding, a 3-gram language model (LM) with a 10K lexicon was used. It results from the interpolation of a generic LM (weight 10%) and a specialized LM (weight 90%). The generic LM was estimated on about 1000M of words from the French newspapers Le Monde and Gigaword. The specialized LM was estimated from the sentences (about 200 words) that the 21 participants had to utter during the experiment.

The Speeral choice was made based on experiments we undertook with several state-of-the-art ASR systems and on the fact that the Driven Decoding Algorithm (DDA) is only implemented in Speeral.

2.4 Driven Decoding Algorithm

DDA aims at aligning and correcting auxiliary transcripts by using a speech recognition engine [25,26]. This algorithm improves system performance dramatically by taking advantage of the availability of the auxiliary transcripts. DDA acts at each new generated assumption of the ASR system. The current ASR assumption is aligned with the auxiliary transcript (from a previous decoding pass). Then a matching score α is computed and integrated with the language model [26].

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 and DDA merges truly and easily the information from the two streams while voting strategies (such as ROVER) do not merge ASR systems outputs ; the applied strategy is dynamic and used, for each utterance to decode, the best SNR channel for the first pass and the second best channel for the last pass. A nice feature of DDA is that it is not impacted by an asynchronous signal segmentation since it works at the word level.

The proposed approach benefits from the multiple microphones of the smart home and from *a priori* knowledge about the sentences being uttered. This approach is based on the DDA which drives an audio stream being decoded by the results of the decoding on another one [10]. The first stream (channel with the best Signal to Noise Ratio) is used to drive the second stream and to improve the

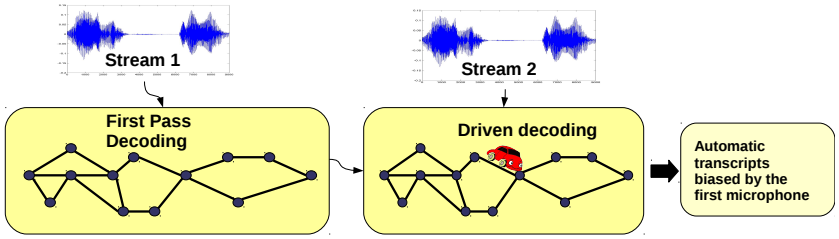

Fig. 3. DDA used with two streams : the first stream drives the second stream

decoding performances by taking into account 2 simultaneous channels (Figure 3). An important aspect to mention is that the purpose of the experiment is to assess the impact of the automatic segmentation. Unlike our previous experiments [10] we do not use grammar in order to bias strongly the ASR system.

3 Multimodal Data Corpus

To provide data to test and train the different processing stages of the Sweet-Home system, experiments were run in the Domus smart home that was designed by the Multicom team of the laboratory of Informatics of Grenoble to observe users' activities interacting with the ambient intelligence of the environment. Figure 5 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 system (KoNneX), a worldwide ISO standard (ISO/IEC 14543) for home and building control. A residential gateway architecture has been designed, supported by a virtual KNX layer seen as an OSGI service (Open Services Gateway Initiative) to guarantee the interoperability of the data coming and to allow the communication with virtual applications, such as activity tracking.

The following sensors were used for multimodal data acquisition :

- 7 radio microphones set into the ceiling that can be recorded in real-time thanks to a dedicated PC embedding an 8-channel input audio card ;
- 3 contact sensors on the furniture doors (cupboards in the kitchen, fridge and bathroom cabinet) ;
- 4 contact sensors on the 4 indoor doors ;
- 4 contact sensors on the 6 windows (open/close) ;
- 2 Presence Infrared Detectors (PID) set on the ceiling.

The multimodal corpus was acquired with 21 persons (including 7 women) acting in the Domus smart home. To make sure that the data acquired would be as close as possible to real daily living data, the participants were asked to perform several daily living activities in the smart home. The average age of the

Fig. 4. Images captured by the DOMUS video cameras during the experiment

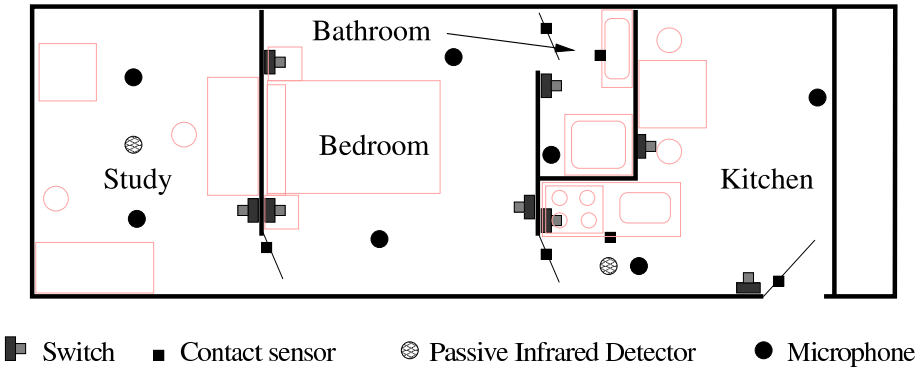

Fig. 5. Layout of the DOMUS Smart Home and position of the sensors.

participants was 38.5 ± 13 years (22-63, min-max). The experiment consisted in following a scenario of activities without condition on the time spent and the manner of achieving them (e.g., having a breakfast, simulate a shower, get some sleep, clean up the flat using the vacuum, etc.). Figure 4 shows participants performing activities in the different rooms of the smart home. A visit, before the experiment, was organized to make sure that the participants will find all the items necessary to perform the activities. During the experiment, event traces from the home automation network, audio and video sensors were captured. Video data were only captured for manual marking up and are not intended to be used for automatic processing. In total, more than 26 hours of data have been acquired.

For the experiment described in this article, we used only the streaming records of the 7 microphones (the remaining data are used for others research work). These records contain the daily living sounds resulting from routine activities during the performance of the scenario as well as two telephone conversations (20 short sentences each in French : “Allô oui”, “C’est moi”, “J’ai mal à la tête” ...).

The human annotation of the audio corpus was a very fastidious task given the high number of sound events generated by the participants. To speed up the process, a detection algorithm was applied to the seven channels to detect intervals of sounds of interest. Then, for human annotation purpose, a unique signal resulting of the combination of the seven channels using a weighted sum

Tab. 1. Detection sensitivity for different values of the overlapping criteria τ.

τ (%)	20	50	80	100
Sensitivity (%)	96.1	93.4	90.3	88.6

with coefficients varying with the signals energy was created. Moreover, sound intervals were fused by making the union of overlapping intervals of the seven channels. This signal, the merged intervals, and the videos were then used by the authors to annotate the sound events. The resulting annotation file contains the start, the end and the class of the sound or speech events.

4 Experimental Results

To assess the performance of the method, audio records from five participants, S01, S03, S07, S09 and S13, were extracted from the multimodal corpus for Sound/Speech discrimination which last respectively 88, 50, 72, 63 and 60 minutes. This small amount of data is due to delay in the annotation process. Furthermore, for sound recognition, S13 was not used due to incomplete annotation. For each subject, the 7 channels were used for detection and classification. In this section, we present the results for the four stages of the system namely : Sound Detection, Sound/Speech Discrimination, Sound Classification and ASR.

4.1 Sound Detection and Extraction

Evaluating the performance in detecting temporal intervals of event is known to be a hard task as several strategies can be employed (discrepancy in start/end occurrence time, duration difference, intersection, etc.). In our case, we used the temporal intersection between the humanly annotated reference intervals and the automatically detected ones. In this strategy, a reference sound event is correctly detected if there is at least τ% overlap between a detected interval and the reference interval. An evaluation of the detection results was made on each channel and a reference interval was considered detected if at least one detection was correct on one of the 7 channels. For τ we tested values between 20% and 100% in order to measure the decrease of sensitivity (or recall). The average results for four person are presented in Table 1. The decrease of sensitivity between $\tau = 20\%$ and $\tau = 50\%$ is less than 3%. In reality if the detected segment contains half of the useful signal, it is sufficient for the sound classification system. This is why the 50% value for τ was chosen.

The detection algorithm was applied to each of the 7 channels in order to be able to make information fusion in the recognition stages. The evaluations were performed with 4 different records of the corpus. The results of the detection ratio are presented in Table 2. The detection sensitivity is very stable across the participants. Moreover, the best detection is also stable over the channel. Indeed, channel 1 and channel 2 gave the best detection sensitivity and exhibited the highest SNR. This is due to the scenario of the corpus which led the participant to be mostly close to these microphones when performing noisy activities.

Tab. 2. Detection Sensitivity for four participants with $\tau = 50\%$.

Participant	S01	S03	S07	S09
Detection Sensitivity (%)	93.2	93.1	93.0	94.4
Average Sensitivity (%)	93.4			

4.2 Sound/Speech Discrimination

For Sound/Speech discrimination, an UBM model of 1024 components was created using 16 MFCC (Mel Frequency Cepstral Coefficients [27]) feature vectors extracted from 16ms signal frames with an overlap of 8ms. The UBM model was learned from speech data from three participants different from those used for the evaluation. The utterances made by the five participants mentioned above were used to adapt the UBM model and generate the supervectors for the SVM stage.

Sound/Speech discrimination was performed on each of the 7 channels. As with the detection, channel 6 and channel 7 gave the best results. Table 3 shows Sound/Speech discrimination results. Number of False Negatives (missed detection or classification of speech) and True Positives (correct classification of speech) are given for detection and Sound/Speech discrimination. The missed utterances were either caused by the detection step (utterance not detected), or by the discrimination step (utterance detected but not recognized as speech). It should be noted that these results do not include false positive recognitions (non-speech sounds recognized as speech, which were actually rare), nor do they take into account speech from radio or improvised phrases that are not be used for speech recognition.

Tab. 3. Sound/Speech discrimination performances.

Participant	# of Utt.	Channel	False Neg. Det.	False Neg. Reco.	True Positive
S01	44	C6	0	4	40
		C7	1	2	41
S03	41	C6	0	2	39
		C7	1	7	33
S07	45	C6	4	5	36
		C7	6	3	36
S09	40	C6	0	0	40
		C7	0	0	40
S13	42	C6	0	4	38
		C7	1	0	41

4.3 Sound Classification

Among 30 annotated sound classes, 18 classes were retained for our experiment (Brushing Teeth, Coughing, Hands Clapping, Door Clapping, Door

Opening, Electric Door Lock, Window Shutters, Curtains, Window Shutters + Curtains, Vacuum Cleaner, Phone Ring, Music Radio, Speech Radio, Speech + Music Radio, Paper, Keys, Kitchenware and Water). Examples of classes not used include typing on a keyboard, eating biscuits, manipulating a plastic bag etc. Although it would have been better to use sounds related to abnormal situations such as human screams or glass breaking, the participants were not asked to perform these kinds of activity. Indeed, the corpus acquisition was performed mainly to test the daily living usage rather than distress situations which are very difficult to produce. Many sounds were either considered as noise or very hard to recognize even by human ears and were annotated as *unknown*.

Given that detection and classification were separately applied to each channel, this led us to a problem of synchronization. Indeed, temporal intervals from a same event recorded on several channels may not be recognized as the same sound class. Our multi-channel aggregation strategy was the following : if an automatic detection does not cover at least 50% of an annotation, then its recognition result will not be taken into account for the classification. To take the final recognition decision using several channels, the sound event with the best SNR is compared to the annotation to compute the classification score.

Table 4 shows the results obtained for each participant. Sounds of Interest (SOI) are the acoustic events in the annotated recordings that belong to the set of the 18 sound classes mentioned above. The measure used to evaluate the performance was the ratio of the number of well recognized detections to the number of detected Sounds of Interest. Results are presented for the detection (column "TP D") and the classification (column "TP C", that is the ratio of well classified SOI taking into account only the number of well detected SOI) as well as for both (column "TP D+C", the ratio of well classified SOI on the total number of SOI). In order to evaluate how far the method is from the optimal performance, the Oracle values are given. The Oracle performances were computed by considering that a SOI is well recognized if at least one of the channels is correct regardless of its SNR. Table 5 shows the average performance per channel for the four subjects. In this experiment, we are only interested in true positives. In other words, for the time being, the system does not include any rejection for unknown sounds. This will be implemented in future work via the use of thresholds, or the creation of one large class for unknown sounds.

Tab. 4. Sound classification performance using the multi-channel fusion strategy.

Subject	# SOI Occur.	TP D(%)	TP C(%)	TP D+C(%)	Oracle TP D+C(%)
S01	230	83.9	69.4	58.2	58.2
S03	175	80.6	65.2	52.6	52.6
S07	245	82.4	68.8	56.7	56.7
S09	268	91.8	74.8	68.7	68.7

Tab. 5. Average performance per channel for all subjects.

Channel	C1	C2	C3	C4	C5	C6	C7	Fusion of all Channels
Avg. TP D+C(%)	31.3	33.3	14.9	24.1	21.8	13.3	9.6	59.1

4.4 Automatic Speech Recognition

The ASR stage was assessed using the classical Word Error Rate (WER) metric : $\frac{insertion+deletion+substitution}{Numberofreferencewords}$ (WER is above 100 if there is more word insertion than reference words). The errors of automatic segmentation (WER ranges between 35.4% to 140%) results in two types of degradation : 1) The insertion of false positive speech detection ; and 2) The deletion of speech that was missed by the detection/discrimination.

In our experiments, the type of degradation differs for each speaker. In order to show the impact of insertions we present the "Global WER" and the "speaker WER". The first one takes into account all the insertions computed out of the reference segmentation while the second one compare the WER only on well detected segments. In all the experiments, the insertions generate a lot of errors and there is constantly a degradation of the "speaker WER" (non detected speech). Nevertheless DDA allows one to improve the WER by taking advantage of the combination of two segmented streams.

Table 6 shows experiments for the ASR system by using the automatic segmentation. We present three baselines ("ref", "ref-DDA" and "mix"). The baselines "ref" and "ref-DDA" are focused on the distant speech recognition issue. On the two "ref" baselines the ASR system is launched on the reference segmentation. The purpose of these baselines is to highlight the encountered difficulties of the ASR system in distant conditions :

- The "ref" baseline is a classical ASR system running on the best SNR channel.
- The "ref-DDA" baseline uses DDA in order to combine the two best SNR channels : this method allows one to improve the ASR robustness.

By using the reference segmentation, the WER ranges between 17.5% and 36.3%. This error rate is mostly due to the distance and the noisy environment. However these results are quite usable for the detection of distress or home automation orders [10] : our previous work using a grammar (with DDA) in order to constrain the ASR system allows to use these high WER rates. Moreover DDA improves by 5% relative of the WER.

In a second time we present the results on the whole system : speech detection, speech segmentation and speech transcription :

- "mix" is a classical ASR system running on the automatic segmentation (and the mix of all channels).
- "DDA" corresponds to the ASR system used in real conditions : DDA acts on the automatic segmentation by combining the two best SNR channels.

The most important WER to consider is related to the speaker : the decision system will be able to filter out false detections. In the presented experiments

this WER ranges between 23% and 54%. Our previous work has shown that despite imperfect recognition the system is still able to detect a high number of original utterances that were actually home automation orders and distress sentences. This is due to the restricted language model coupled with the DDA strategy which permits to retrieve correct sentences that are of lowest likelihood than the first hypothesis among the set of hypotheses [28]

Tab. 6. Speech recognition results for each speaker.

	Correct Words (%)	Global WER (%)	speaker WER (%)
S01-ref	84.4	17.5	17.5
S01-ref-DDA	84.0	16.0	16.0
S01-mix	77.4	36.8	24.1
S01-DDA	78.7	35.4	23.1
S03-ref	87.3	20.1	20.1
S03-ref-DDA	86.0	20.9	20.9
S03-mix	32.3	105.8	75.7
S03-DDA	56.0	74.1	51.9
S07-ref	69.6	36.3	36.3
S07-ref-DDA	69.7	36.2	36.2
S07-mix	47.9	81.8	56.5
S07-DDA	50.2	81.8	54.0
S09-ref	85.1	15.4	15.4
S09-ref-DDA	85.3	15.3	15.3
S09-mix	69.1	57.4	32.4
S09-DDA	69.1	54.3	31.9
S13-ref	75.3	27.4	27.4
S13-ref-DDA	76.4	26.3	26.3
S13-mix	44.7	140.5	57.4
S13-DDA	47.9	120.0	54.5

5 Discussion and conclusion

In this paper we propose a complete system for sound analysis in smart home. The system is designed hierarchically and can deal with multiple channels. This multi-layer design makes it easy to test the performance of each module separately. The sound detection module detects most sounds events in the house thanks to the use of 7 channels. Sound/Speech discrimination gave good result for the two channels with highest SNR as the subject does not move around the house whilst talking. Sound classification was more challenging because of the greater number of sound classes and the fact that the "best" channel(s) varies all the the time and some kind of sound events can occur anywhere in the house. Multi-channel fusion strategy based on SNR gave very encouraging results (Table 4 and 5), taking into account that the sound detection's performance also affects

that of sound classification. However, the current system does not include any rejection. We intend to implement this feature in future work. Furthermore, as many daily sounds vary considerably in terms of representation, it would be very desirable to be able to use simpler features for sound classification, at least for some sounds. This will also be investigated in further work.

As for the ASR system, the proposed approach based on DDA lead to moderate robustness. The impact of automatic segmentation for the ASR system highlights new challenges for integration within a smart home ; each stage spreading its errors. Despite the occasionally low performance, the ASR system offers the possibility to be exploited in industrial context. Our future work will focus on the cooperation with the decision-making module.

Références

1. Chan, M., Campo, E., Estève, D., Fourniols, J.Y. : Smart homes — current features and future perspectives. Maturitas **64**(2) (2009) 90–97
2. Weiser, M. : The computer for the 21st century. Scientific American **265**(3) (1991) 66–75
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. 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
5. 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
6. 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
7. Badii, A., Boudy, J. : CompanionAble - integrated cognitive assistive & domotic companion robotic systems for ability & security. In : 1st Congres of the Société Française des Technologies pour l'Autonomie et de Gérontechnologie (SFTAG'09), Troyes (2009) 18–20
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** (2009)
9. Filho, G., Moir, T.J. : From science fiction to science fact : a smart-house interface using speech technology and a photo-realistic avatar. International Journal of Computer Applications in Technology **39**(8) (2010) 32–39
10. Lecouteux, B., Vacher, M., Portet, F. : Distant Speech Recognition in a Smart Home : Comparison of Several Multisource ASRs in Realistic Conditions. In : Interspeech 2011, Florence, Italy (aug 2011) 4p.
11. Chen, J., Kam, A.H., Zhang, J., Liu, N., Shue, L. : Bathroom activity monitoring based on sound. In : Pervasive'05. (2005) 47–61

12. 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 (in press)
13. 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
14. Jaakkola, T., Haussler, D. : Exploiting generative models in discriminative classifiers. In : In Advances in Neural Information Processing Systems 11, MIT Press (1998) 487–493
15. Temko, A., Monte, E., Nadeu, C. : Comparison of sequence discriminant support vector machines for acoustic event classification. In : In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. (2005)
16. Wan, V., Renals, S. : Speaker verification using sequence discriminant support vector machines. IEEE Transactions on Speech and Audio Processing (2005) 203–210
17. Campbell, W.M., Sturim, D.E., Reynolds, D.A., Solomonoff, A. : SVM based speaker verification using a gmm supervector kernel and nap variability compensation. In : in Proceedings of ICASSP, 2006. (2006) 97–100
18. Fauve, B., Matrouf, D., Scheffer, N., Bonastre, J.F. : State-of-the-art performance in text-independent speaker verification through open-source software. In : IEEE Transactions on Audio, Speech, and Language Processing. Volume 15. (2007) 1960–1968
19. Burges, C.J.C. : A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. (1998) 121–167
20. Schölkopf, B., Smola, A.J. : Learning with Kernels. MIT Press (2002)
21. Sehili, M.A., Istrate, D., Boudy, J. : Primary investigations of sound recognition for a domotic application using support vector. Annals of the University of Craiova, Series : Automation, Computers, Electronics and Mechatronics **7(34)**(2) (2010) 61–65
22. Reynolds, D.A., Quatieri, T.F., Dunn, R.B. : Speaker verification using adapted gaussian mixture models. In : Digital Signal Processing. (2000) 2000
23. Wölfel, M., McDonough, J. : Distant Speech Recognition. John Wiley and Sons, 573 pages (2009)
24. Linarès, G., Nocéra, P., Massonié, D., Matrouf, D. : The LIA speech recognition system : from 10xRT to 1xRT. In : Proc. TSD'07. (2007) 302–308
25. 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
26. Lecouteux, B., Linarès, G., Bonastre, J., Nocéra, P. : Imperfect transcript driven speech recognition. In : InterSpeech'06. (2006) 1626–1629
27. Logan, B. : Mel frequency cepstral coefficients for music modeling. In : Proceedings of International Symposium on Music Information Retrieval. (2000)
28. Vacher, M., Lecouteux, B., Portet, F. : Recognition of Voice Commands by Multisource ASR and Noise Cancellation in a Smart Home Environment. In : EUSIPCO, Bucarest, Romania (Aug 2012) 1663–1667