Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living

Michel Vacher[1], Anthony Fleury[2], François Portet[1], Jean-François Serignat[1], Norbert Noury[2]
[1]*Laboratoire d'Informatique de Grenoble, GETALP team, Université de Grenoble France*
[2]*Laboratory TIMC-IMAG, AFIRM team, Université de Grenoble France*

1. Introduction

Recent advances in technology have made possible the emergence of Health Smart Homes (Chan et al., 2008) designed to improve daily living conditions and independence for the population with loss of autonomy. Health smart homes are aiming at assisting disabled and the growing number of elderly people which, according to the World Health Organization (WHO), is forecasted to reach 2 billion by 2050. Of course, one of the first wishes of this population is to be able to live independently as long as possible for a better comfort and to age well. Independent living also reduces the cost to society of supporting people who have lost some autonomy. Nowadays, when somebody is loosing autonomy, according to the health system of her country, she is transferred to a care institution which will provide all the necessary supports. Autonomy assessment is usually performed by geriatricians, using the index of independence in Activities of Daily Living (ADL) (Katz & Akpom, 1976), which evaluates the person's ability to realize different activities of daily living (e.g., doing a meal, washing, going to the toilets . . .) either alone, or with a little or total assistance. For example, the AGGIR grid (*Autonomie Gérontologie Groupes Iso-Ressources*) is used by the French health system. Seventeen activities including ten discriminative (e.g., talking coherently, orientating himself, dressing, going to the toilets...) and seven illustrative (e.g., transports, money management, ...) are graded with an A (the task can be achieved alone, completely and correctly), a B (the task has not been totally performed without assistance or not completely or not correctly) or a C (the task has not been achieved). Using these grades, a score is computed and, according to the scale, a geriatrician can deduce the person's level of autonomy to evaluate the need for medical or financial support.

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 harmful situations (e.g., fall, gas not turned off). Basically, an health smart home contains sensors used to monitor the activity of the inhabitant. The sensors data is analyzed to detect the current situation and to execute the appropriate feedback or assistance. One of the first steps to achieve these goals is to detect the daily activities and to assess the evolution of the

APPLE 1051

monitored person's autonomy. Therefore, activity recognition is an active research area (Albinali et al., 2007; Dalal et al., 2005; Duchêne et al., 2007; Duong et al., 2009; Fleury, 2008; Moore & Essa, 2002) but, despite this, it has still not reached a satisfactory performance nor led to a standard methodology. One reason is the high number of flat configurations and available sensors (e.g., infra-red sensors, contact doors, video cameras, RFID tags, etc.) which may not provide the necessary information for a robust identification of ADL. Furthermore, to reduce the cost of such an equipment and to enable interaction (i.e., assistance) the chosen sensors should serve not only to monitor but also to provide feedback and to permit direct orders.
One of the modalities of choice is the audio channel. Indeed, audio processing can give information about the different sounds in the home (e.g., object falling, washing machine spinning, door opening, foot step ...) but also about the sentences that have been uttered (e.g., distress situations, voice commands). Moreover, speaking is the most natural way for communication. A person, who cannot move after a fall but being concious has still the possibility to call for assistance while a remote controller may be unreachable.
In this chapter, we present AUDITHIS— a system that performs real-time sound and speech analysis from eight microphone channels — and its evaluation in different settings and experimental conditions. Before presenting the system, some background about health smart home projects and the *Habitat Intelligent pour la Santé* of Grenoble is given in section 2. The related work in the domain of sound and speech processing in Smart Home is introduced in section 3. The architecture of the AUDITHIS system is then detailed in section 4. Two experimentations performed in the field to validate the detection of distress keywords and the noise suppression are then summarised in section 5. AUDITHIS has been used in conjunction with other sensors to identify seven Activities of Daily Living. To determine the usefulness of the audio information for ADL recognition, a method based on feature selection techniques is presented in section 6. The evaluation has been performed on data recorded in the Health Smart Home of Grenoble. Both data and evaluation are detailed in section 7. Finally, the limits and the challenges of the approach in light of the evaluation results are discussed in section 8.

2. Background

Health smart homes have been designed to provide ambient assisted living. This topic is supported by many research programs around the world because ambient assisted living is supposed to be one of the many ways to aid the growing number of people with loss of autonomy (e.g., weak elderly people, disabled people ...). Apart from supporting daily living, health smart homes constitute a new market to provide services (e.g., video-conferencing, tele-medicine, etc.). This explains the involvement of the major telecommunication companies. Despite these efforts, health smart home is still in its early age and the domain is far from being standardised (Chan et al., 2008). In the following section, the main projects in this field — focusing on the activity recognition — are introduced. The reader is referred to (Chan et al., 2008) for an extensive overview of smart home projects. The second section is devoted to the Health Smart Home of the TIMC-IMAG laboratory which served for the experiments described further in this chapter.

2.1 Related Health Smart Home Projects

To be able to provide assistance, health smart homes need to perceive the environment — through sensors — and to infer the current situation. Recognition of activities and distress situations are generally done by analyzing the evolution of indicators extracted from the sensors raw signals. A popular trend is to use as many as possible sensors to acquire the most

information. An opposite direction is to use the least number of sensors as possible to reduce the cost of the smart home. For instance, the Edelia company[1] evaluates the quantity of water used per day. A model is built from these measurements and in case of high discrepancy between the current water use and the model, an alert to the relatives of the inhabitant is generated. Similar work has been launched by Zojirushi Corporation[2] which keeps track of the use of the electric water boiler to help people stay healthy by drinking tea (which is of particular importance in Japan). In an hospital environment, the Elite Care project (Adami et al., 2003) proposed to detect the bedtime and wake-up hours to adapt the care of patients with Alzheimer's disease.

These projects focus on only one sensor indicator but most of the research projects includes several sensors to estimate the '*model*' of the lifestyle of the person. The model is generally estimated by data mining techniques and permits decision being made from multisource data. Such smart homes are numerous. For instance, the project *House_n* from the Massachusetts Institute of Technology, includes a flat equipped with hundreds of sensors (Intille, 2002). These sensors are used to help performing the activities of daily living, to test Human-Machine Interfaces, to test environment controller or to help people staying physically and mentally active. This environment has been designed to easily assess the interest of new sensors (e.g., RFID, video camera, etc.). A notable project, *The Aware Home Research Initiative* (Abowd et al., 2002) by the Georgia Institute of Technology, consists in a two-floor home. The ground floor is devoted to an elderly person who lives in an independent manner whereas the upper floor is dedicated to her family. This family is composed of a children mentally disabled and his parents who raise him while they work full-time. This house is equipped with motion and environmental sensors, video cameras (for fall detection and activity recognition (Moore & Essa, 2002) and short-term memory help (Tran & Mynatt, 2003)) and finally RFID tags to find lost items easily. Both floors are connected with flat screens to permit the communication of the two generations. The AILISA (LeBellego et al., 2006) and PROSAFE (Bonhomme et al., 2008) projects have monitored the activities of the person with presence infra-red sensors to raise alarms in case of abnormal situations (e.g., changes in the level of activities). Within the PROSAFE project, the ERGDOM system controls the comfort of the person inside the flat (i.e., temperature, light...).

Regarding the activity detection, although most of the many researches related to health smart homes is focused on sensors, network and data sharing (Chan et al., 2008), a fair number of laboratories started to work on reliable Activities of Daily Living (ADL) detection and classification using Bayesian (Dalal et al., 2005), rule-based (Duong et al., 2009; Moore & Essa, 2002), evidential fusion (Hong et al., 2008), Markovian (Albinali et al., 2007; Kröse et al., 2008), Support Vector Machine (Fleury, 2008), or ensemble of classifiers (Albinali et al., 2007) approaches. For instance, (Kröse et al., 2008) learned models to recognize two activities: '*going to the toilets*' and '*exit from the flat*'. (Hong et al., 2008) tagged the entire fridge content and other equipments in the flat to differentiate the activities of preparing cold or hot drinks from hygiene. Most of these approaches have used Infra-red sensors, contact doors, videos, RFID tags etc. But, to the best of our knowledge, only few studies include audio sensors (Intille, 2002) and even less have assessed what the important features (i.e. sensors) for robust classification of activities are (Albinali et al., 2007; Dalal et al., 2005). Moreover, these projects considered only few activities while many daily living activities detection is required for autonomy assessment. Our approach was to identify seven activities of daily living that will be useful for

[1] www.edelia.fr/
[2] www.zojirushi-world.com/

the automatic evaluation of autonomy, and then to equip our Health Smart Home with the most relevant sensors to learn models of the different activities (Portet et al., 2009). The next section details the configuration of health smart home.

2.2 The TIMC-IMAG's Health Smart Home

Since 1999, the TIMC-IMAG laboratory in Grenoble set-up, inside the faculty of medicine of Grenoble, a flat of 47m^2 equipped with sensing technology. This flat is called *HIS* from the French denomination: *Habitat Intelligent pour la Santé* (i.e., Health Smart Home). The sensors and the flat organization are presented in Figure 1. It includes a bedroom, a living-room, a corridor, a kitchen (with cupboards, fridge...), a bathroom with a shower and a cabinet. It has been firstly equipped with presence infra-red sensors, in the context of the AILISA project (LeBellego et al., 2006) and served as prototype for implementation into two flats of elderly persons and into hospital suites of elderly people in France. Important features brought by the infra-red sensors have been identified such as mobility and agitation (Noury et al., 2006) (respectively the number of transitions between sensors and the number of consecutive detections on one sensor) which are related to the health status of the person (Noury et al., 2008). The HIS equipment has been further complemented with several sensors to include:

- *presence infra-red sensors* (PIR), placed in each room to sense the location of the person in the flat;
- *door contacts*, for the recording of the use of some furniture (fridge, cupboard and dresser);
- *microphones*, set in each room to process sounds and speech; and
- *large angle webcams*, that are placed only for annotation purpose.

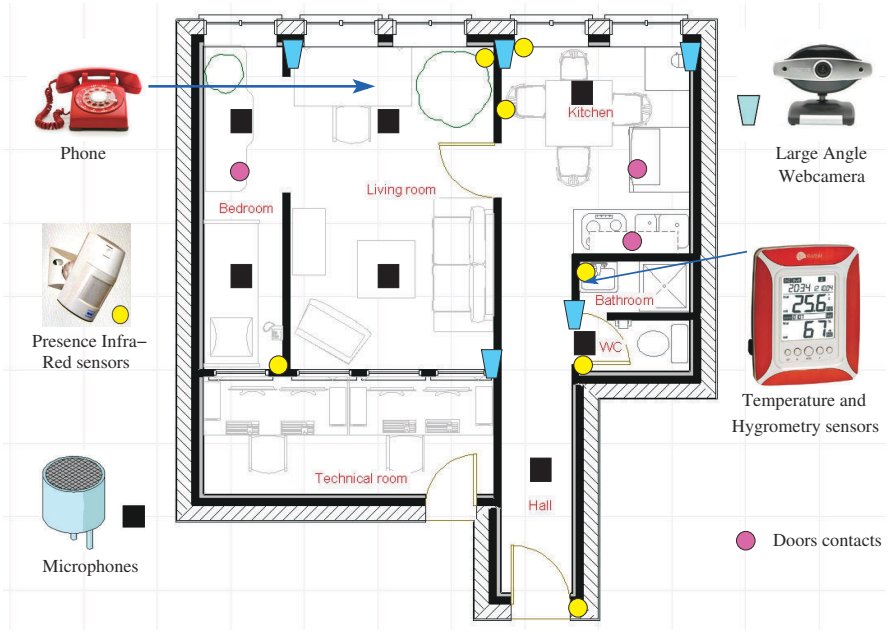

Fig. 1. The Health Smart Home of the TIMC-IMAG Laboratory in Grenoble

The cost of deployment of such installation is reduced by using only the sensors that are the most informative. This explains the small number of sensors compared to other smart homes (Intille, 2002). The technical room contains 4 standard computers which receive and store, in real time, the information from the sensors. The sensors are connected with serial port (contact-doors), USB port (webcams), wireless receiver (PIRs) or through an analog acquisition board (microphones). Except for the microphones these connections are available on every (even low-cost) computer. These sensors were chosen to enable the recognition of activities of daily living, such as sleeping, preparing and having a breakfast dressing and undressing, resting, etc. The information that can be extracted from these sensors and the activities they are related to are summarised in Table 5 presented in section 7.

It is important to note that this flat represents an hostile environment for information acquisition similar to the one that can be encountered in real home. This is particularly true for the audio information. For example, we have no control on the sounds that are measured from the exterior (e.g., the flat is near the helicopter landing strip of the local hospital). Moreover, there is a lot of reverberation because of the 2 important glazed areas opposite to each other in the living room. The sound and speech recognition system presented in section 4 has been tested in laboratory and gave an average Signal to Noise Ratio of 27dB in-lab. In the HIS, this fell to 12dB. Thus, the signal processing and learning methods that are presented in the next sections have to address the challenges of activity recognition in such a noisy environment.

3. State of the Art in the Context of Sound and Speech Analysis

Automatic sound and speech analysis are involved in numerous fields of investigation due to an increasing interest for automatic monitoring systems. Sounds can be speech, music, songs or more generally sounds of the everyday life (e.g., dishes, step,...). This state of the art presents firstly the sound and speech recognition domains and then details the main applications of sound and speech recognition in smart home context.

3.1 Sound Recognition

Sound recognition is a challenge that has been explored for many years using machine learning methods with different techniques (e.g., neural networks, learning vector quantizations,...) and with different features extracted depending on the technique (Cowling & Sitte, 2003). It can be used for many applications inside the home, such as the quantification of water use (Ibarz et al., 2008) but it is mostly used for the detection of distress situations. For instance, (Litvak et al., 2008) used microphones to detect a special distress situation: the fall. An accelerometer and a microphone are both placed on the floor. Mixing sound and vibration of the floor allowed to detect fall of the occupant of the room. (Popescu et al., 2008) used two microphones for the same purpose, using Kohonen Neural Networks. Out of a context of distress situation detection, (Chen et al., 2005) used HMM with the Mel-Frequency Cepstral Coefficients (MFCC) to determine the different uses of the bathroom (in order to recognize sequences of daily living). (Cowling, 2004) applied the recognition of non-speech sounds associated with their direction, with the purpose of using these techniques in an autonomous mobile surveillance robot.

3.2 Speech Recognition

Human communication by voice appears to be so simple that we tend to forget how variable a signal speech is. In fact, spoken utterances even of the same text are characterized by large

differences that depend on context, speaking style, the speaker's dialect, the acoustic environment... Even identical texts spoken by the same speaker can show sizable acoustic differences. Automatic methods of speech recognition must be able to handle this large variability in a fault-free fashion and thus the progress in speech processing are not as fast as hoped at the time of the early work in this field.

The phoneme duration, the fundamental frequency (melody) and the Fourier analysis have been used for studying phonograph recordings of speech in 1906. The concept of short-term representation of speech, where individual feature vectors are computed from short (10-20 ms) semi-stationary segments of the signal, were introduced during the Second World War. This concept led to a spectrographic representation of the speech signal and to underline the importance of the formants as carriers of linguistic information. The first recognizer used a resonator tuned to the vicinity of the first formant vowel region to trigger an action when a loud sound were pronounced. This knowledge-based approach were abandoned by the first spoken digit recognizer in 1952 (Davis et al., 1952). (Rabiner & Luang, 1996) published the scaling algorithm for the Forward-Backward method of training of Hidden Markov Model recognizers and at this time modern general-purpose speech recognition systems are generally based on HMMs as far as the phonemes are concerned. Models of the targeted language are often used. A Language model is a collection of constraints on the sequence of words acceptable on a given language and may be adapted to a particular application. The specificities of a recognizer are related to its adaptation to a unique speaker or to a large variety of speakers, and to its capacities of accepting continuous speech, and small or large vocabularies. Many computer softwares are nowadays able to transcript documents on a computer from speech that is uttered at normal pace (for the person) and at normal loud in front of a microphone connected to the computer. This technique necessitates a learning phase to adapt the acoustic models to the person. That is done from a given set of sentences uttered by the speaker the first time he used the system. Dictation systems are capable of accepting very large vocabularies, more than ten thousand words. Another kind of application aims to recognize a small set of commands, i.e. for home automation purpose or on a vocal server (of an answering machine for instance). This can be done without a speaker adapted learning step (that would be too complicated to set-up). Document transcription and command recognition use speech recognition but have to face different problems in their implementation. The first application needs to be able to recognize, with the smallest number of mistakes, a large number of words. For the second application, the number of words is lower, but the conditions are worst. Indeed, the use of speech recognition to enter a text on a computer will be done with a good microphone, well placed (because often associated to the headphone) and with relatively stable conditions of noise on the measured signal. In the second application, the microphone could be, for instance, the one of a cell phone, that will be associated to a low-pass filter to reduce the transmissions on the network, and the use could be done in every possible conditions (e.g., in a train with a baby crying next to the person).

More general applications are for example related to the context of civil safety. (Clavel et al., 2007) studied the detection and analysis of abnormal situations through fear-type acoustic manifestations. Two kinds of application will be presented in the continuation of this section: the first one is related to people aids and the second one to home automation.

3.3 Speech and Sound Recognition Applied to People Aids

Speech and sound recognition have been applied to the assistance to the person. For example, based on a low number of words, France Telecom Research and Development worked on a

pervasive scarf that can be useful to elderly or dependant people (with physical disabilities for instance) in case of problem. It allows to call, easily (with vocal or tactile commands) a given person (previously registered) or the emergencies.
Concerning disabled or elderly people, (Fezari & Bousbia-Salah, 2007) have demonstrated the feasibility to control a wheel chair using a given set of vocal commands. This kind of commands uses existing speech recognition engines adapted to the application. In the same way, Renouard et al. (2003) worked on a system with few commands able to adapt continuously to the voice of the person. This system is equipped with a memory that allows the training of a reject class.
Finally, speech recognition can be used to facilitate elderly people access to new technologies. For example, Kumiko et al. (2004) aims at assisting elderly people that are not familiar with keyboards through the use of vocal commands. Anderson et al. (1999) proposed the speech recognition of elderly people in the context of information retrieval in document databases.

3.4 Application of Speech and Sound Recognition in Smart Homes

Such recognition of speech and sound can be integrated into the home for two applications:

- Home automation,
- Recognition of distress situations.

For home automation, (Wang et al., 2008) proposed a system based on sound classification, this allows them to help or to automatize tasks in the flat. This system is based on a set of microphones integrated into the ceiling. Classification is done with Support Vector Machines from the MFCC coefficients of the sounds.
Recognition of distress situations may be achieved through sound or speech analysis; a distress situation being recognized when some distress sentences or key words are uttered, or when some sounds are emitted in the flat like glass breaking, screams or object falling. This was explored by (Maunder et al., 2008) which constructed a database of sounds of daily life acquired by two microphones in a kitchen. They tried to differentiate sounds like phone, dropping a cup, dropping a spoon, etc. using Gaussian Mixture Models. (Harma et al., 2005) collected sounds in an office environment and tried unsupervised algorithms to classify the sounds of daily life at work. Another group, (Istrate et al., 2008), aimed at recognizing the distress situations at home in embedded situations using affordable material (with classical audio sound cards and microphones).
On another direction, researches have been engaged to model the dialogue of an automated system with elderly people (Takahashi et al., 2003). The system performs voice synthesis, speech recognition, and construction of a coherent dialogue with the person. This kind of research have application in robotics, where the aim is then to accompany the person and reduce his loneliness.
Speech and sound analyses are quite challenging because of the recording conditions. Indeed, the microphone is almost never placed near the speaker or embedded, but often set in the ceiling. Surrounding noise and sound reverberation can make the recognition very difficult. Therefore, speech and sound recognition have to face different kind of problems. Thus a signal processing adapted to the recording conditions is requested. Moreover, automatic speech recognition necessitates acoustic models (to identify the different phonemes) and languages models (recognition of words) adapted to the situation. Elderly people tends to have voice characteristics different from the active population (Wilpon & Jacobsen, 1996). (Baba et al., 2004) constructed specifically acoustic models for this target population to asses the usefulness of such adaptation.

Our work consists in a complete sound recognition system to identify the different sounds in the flat in order to recognize the currently performed activity of daily living, associated to a speech recognition system in French to search for distress keywords inside the signal measured. The implementation and test of this complete system is described in the next sections.

4. The AUDITHIS and RAPHAEL Systems

The term AUDITHIS is built from the names audit and audition, and the acronym HIS (*Habitat Intelligent pour la Santé* - Health Smart Home) and the merger of audio and audit, because the system aims at sound and speech analysis in a health smart home. Therefore, AUDITHIS is able to analyze, in real-time, information from eight microphones placed at different location of a smart home. Figure 2 depicts the general organization of the AUDITHIS audio analysis system and its interaction with the Autonomous Speech Recognizer RAPHAEL.

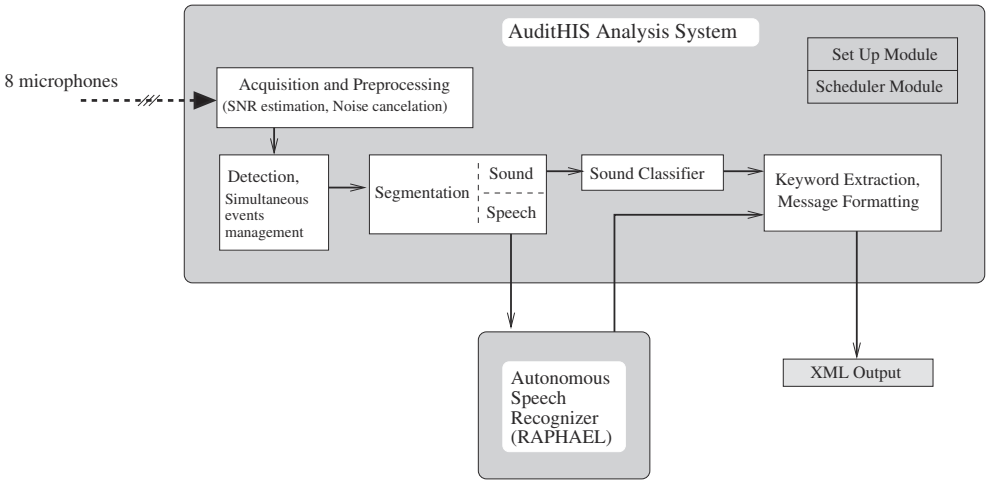

Fig. 2. Architecture of the AUDITHIS and RAPHAEL systems

Both systems are running in real-time as independent applications on the same GNU/Linux operating system and they are synchronized through a file exchange protocol. Each of the 8 microphones is connected to an analog input channel of the acquisition board. Each sound is processed independently and sucessively by the different modules thanks to a queuing management protocol:

1. **Data Acquisition and preprocessing**, which is in charge of signal acquisition, SNR estimation, noise cancellation;
2. **Detection**, which estimates the beginning and end of a sound to analyse and manage the simultaneous audio events;
3. **Segmentation**, which classifies each audio event as being speech or sound of daily living;
4. **Sound classification or Speech Recognition (**RAPHAEL**)**, which determines which class of sound or which phrase has been uttered; and
5. **Message Formatting**.

These modules run as independent threads synchronized by a scheduler. The following sections detail each of the modules.

4.1 Data Acquisition and preprocessing

Data acquisition is operated on the 8 input channels simultaneously at a 16 kHz sampling rate by the first module. Data of each channel is stored in a buffer and processed sequentially and separately. Noise level is also evaluated by this module to assess the Signal to Noise Ratio (SNR) of each acquired sound. The SNR of each audio signal is very important for the decision system to estimate the reliability of the corresponding analysis output. Moreover, noise suppression techniques are incorporated in this module in order to suppress on the fly the noise emitted by known sources like TV or radio; this part of the module is described in section 4.2.

4.2 Known Source Noise Suppression

Listening to the radio and watching TV are very frequent everyday activities; this can seriously disturb a sound and speech recognizer. Because of that, sound and speech analysis must solve two problems: firstly, sounds or speech emitted by the person in the flat can be altered by the loudspeaker and badly recognized, and secondly, radio and TV sounds will be analyzed as well although their information is not relevant. It will be then mandatory to take into account the fact that the radio or the TV is up to suppress this noise or to exploit the resulting information in an other way. Sound $x(n)$ emitted by a loudspeaker in the health smart home is a noise source that will be altered by the room acoustics depending on the position of the microphone in the room. The resulting noise $y(n)$ of this alteration may be expressed by a convolution product in the time domain (Equation 1), h being the impulse response and n the discrete time.

$$y(n) = h(n) * x(n) \tag{1}$$

This noise is then superposed to the interesting signal $e(n)$ emitted in the room: speech uttered by the person or everyday life sound. The signal recorded by the microphone is then $y(n) = e(n) + h(n) * x(n)$. Various methods were developed in order to cancel the noise (Michaut & Bellanger, 2005), some methods attempt to obtain $\widehat{h}(n)$ an estimation of the impulse response of the room in order to remove the noise as shown on Figure 3. The resulting output is given in Equation 2.

$$v(n) = e(n) + y(n) - \widehat{y}(n) = e(n) + h(n) * x(n) - \widehat{h}(n) * x(n) \tag{2}$$

These methods may be divided into 2 classes: Least Mean Square (LMS) and Recursive Least Square (RLS) methods. Stability and convergence properties are studied in (Michaut & Bellanger, 2005). The Multi-delay Block Frequency Domain (MDF) algorithm is an implementation of the LMS algorithm in the frequency domain (Soo & Pang, 1990). In echo cancellation systems, the presence of audio signal $e(n)$ (double-talk) tends to make the adaptive filter diverge. To prevent this problem, robust echo cancellers require adjustment of the learning rate to take the presence of double talk in the signal into account. Most echo cancellation algorithms attempt to explicitly detect double-talk but this approach is not very successful, especially in presence of a stationary background noise. A new method (Valin & Collings, 2007) was proposed by the authors of the library, where the misalignment is estimated in closed-loop based on a gradient adaptive approach; this closed-loop technique is applied to the block frequency domain (MDF) adaptive filter.

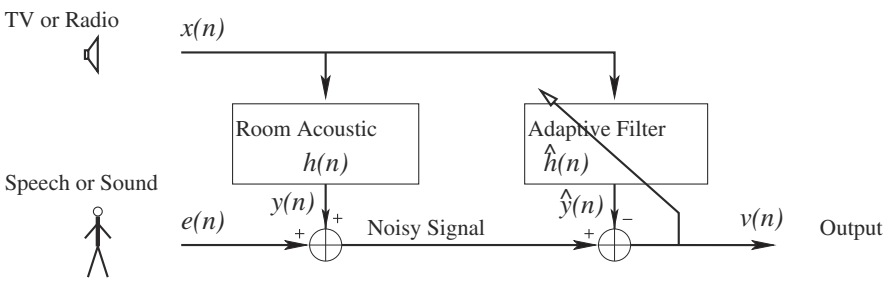

Fig. 3. Echo Cancellation System used for Noise Suppression

The echo cancellation technique used introduces a specific noise into the $v(n)$ signal and a post-filtering is requested. This algorithm is implemented in the SPEEX library under GPL License (Valin, 2007) for echo cancellation system. The method implemented in this library is the Minimum Mean Square Estimator Short-Time Amplitude Spectrum Estimator (MMSE-STSA) presented in (Ephraim & Malah, 1984). The STSA estimator is associated to an estimation of the *a priori* SNR. The formulated hypothesis are following:

- added noise is Gaussian, stationary and the spectral density is known;
- an estimation of the speech spectrum is available;
- spectral coefficients are Gaussian and statistically independents;
- the phase of the Discrete Fourier Transform follows a uniform distribution law and is amplitude independent.

Some improvements are added to the SNR estimation (Cohen & Berdugo, 2001) and a psychoacoustical approach for post-filtering (Gustafsson et al., 2004) is implemented. The purpose of this post-filter is to attenuate both the residual echo remaining after an imperfect echo cancellation and the noise without introducing '*musical noise*' (i.e. randomly distributed, time-variant spectral peaks in the residual noise spectrum as spectral subtraction or Wiener rule does (Vaseghi, 1996)). The post-filter is implemented in the frequency domain, which basically means that the spectrum of the input signal is multiplied by weighting coefficients. Their weighted values are chosen by taking into account auditory masking. Noise is inaudible if it is too close to the useful signal in frequency or time; therefore noise components which lie below the masked threshold of the ear are inaudible and can thus be left unchanged. This method leads to more natural hearing and to less annoying residual noise.

4.3 Detection

The detection module is in charge of signal extraction, i.e. to detect the beginning and the end of the audio event. The first step detects the portion of signal that corresponds to a sound segment. It evaluates the background noise of the room and determines a threshold of detection from this. If this adaptive threshold is exceeded by the energy of wavelet trees of highest order level (3 level depth), the signal of the channel is recorded until its energy becomes lower than a second adaptive threshold. Each event is stored in a file for further analysis by the segmentation and recognition modules. The complete method for the detection of the bounds of a given event and also the associated evaluations is described in (Istrate et al., 2006).

4.4 Segmentation

The segmentation module is a Gaussian Mixture Model (GMM) classifier which classifies each audio event as everyday life sound or speech. The segmentation module was trained with an everyday life sound corpus (Vacher et al., 2007) and with the Normal/Distress speech corpus recorded in our laboratory (Vacher et al., 2008). Acoustical features are Linear-Frequency Cepstral Coefficients (LFCC) with 16 filter banks; the classifier uses 24 Gaussian models. These features are used because life sounds are better discriminated from speech with constant bandwidth filters, than with Mel-Frequency Cepstral Coefficients (MFCC), on a logarithmic Mel scale (Vacher et al., 2007). MFCC are the most widely used features for speech recognition. Acoustical features are evaluated using frames whose width is of 16 ms, with an overlap of 50%.

4.5 Sound Classification

Everyday life sounds are classified with either a GMM or Hidden Markov Model (HMM) classifier; the classifier is chosen at the beginning of the experiment. The models were trained with our corpus containing the eight classes of everyday life sounds, using LFCC features (24 filter banks) and 12 Gaussian models. The sound classes are: dishes sounds, door lock, door slap, glass breaking, object falls, ringing phone, screams and step sounds. This corpus is made of 1985 sounds and its total duration is 35 min 38 s. The HMM classifier gives best results in noiseless conditions but we chose the GMM classifier that gives best results when the SNR is under +10 dB. The models could be extended to include more daily living sounds requested to operate in the real life.

4.6 Speech Recognition: the RAPHAEL ASR

The autonomous speech recognizer RAPHAEL (Vacher et al., 2008) is running as an independent application. It analyzes the speech events resulting from the segmentation module, through a file exchange protocol. As soon as an input file is analyzed, it is deleted, and the 5 best hypothesizes are stored in a file. This event allows the AuditHIS scheduler to send the next queued file to the recognizer. Moreover, each sentence file is stored in order to allow future analysis with different recognition parameters of the recognizer. The architecture of the ASR is described by Figure 4. The first stage is the audio interface in charge of acoustical feature extraction in each 16 ms frame with a 50% overlay. 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* (HMMs), the *phonetic dictionary* and the *language models* (tri-grams). The output of the recognizer is made of the 5 best hypothesis lattices.

The training of the *acoustic models* was made with large corpora in order to ensure good speaker independence. These corpora were recorded by 300 French speakers by our team (BRAF100) (Vaufreydaz et al., 2000) and at the LIMSI laboratory (BREF80 and BREF120) (Gauvain et al., 1990). The phonetic dictionary consists in the association of each word in French with its phoneme sequence using the SAMPA coding. Some phonetic variants were added to take into account the possible liaison between word or a possible incorrect pronunciation (e.g., the confusion between the closed vowel [e] and the open vowel [E]).

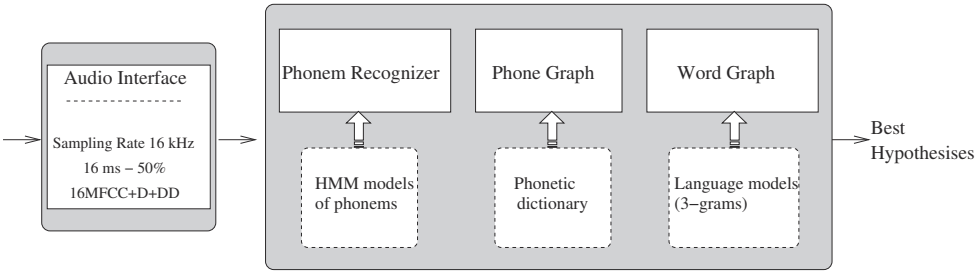

Fig. 4. Architecture of the AUDITHIS and RAPHAEL systems

Sample	Domotic Order	Distress Sentence	Usual Sentence
1	Allume la lumière	A l'aide	Allo c'est moi
2	Eteins la lumière	Je suis tombé	Allo c'est qui
3	Ferme la porte	Une infirmière vite	Bonjour Monsieur
4	Ouvre la porte	Appelez une ambulance	Dehors il pleut
5	Fermez les volets	Aïe aïe aïe	Euh non
6	Ouvrez les volets	Je ne peux plus bouger	J'ai bu du café
7	Il fait très chaud	Je ne me sens pas bien du tout	J'ai fermé la fenêtre
8	Il fait très froid	Je me sens très mal	J'ai sommeil
9	J'ai très chaud	J'ai mal	Tout va bien
10	J'ai très froid	J'ai de la fièvre	A demain

Table 1. Excerpt of the colloquial corpus

The *language model* of this system is a small vocabulary statistical system (299 words in French). The language model is made of 299 uni-grams, 729 bi-grams and 862 trigrams, it is obtained using textual information of a colloquial corpus in French. Our main requirement is the correct detection of a possible distress situation through keyword detection, without understanding the patient's speech. This colloquial corpus contains the sentences in the Normal/Distress speech corpus (Vacher et al., 2006), along with sentences currently uttered during a phone conversation: '*Allo oui*', '*A demain*', '*J'ai bu ma tisane*', '*Au revoir*' etc. and sentences that may be a command for a home automation system. The Normal/Distress language corpus is composed of 126 sentences in French in which 66 are every day sentences: '*Bonjour*' ('*Hello*'), '*Où est le sel?*' ('*Where is the salt?*') ...and 60 are distress phrases: '*Aouh !*', '*Aïe !*', '*Au secours !*' ('*Help !*'), '*Un médecin vite !*' ('*A doctor! hurry!*') along with incorrect grammatically phrases such as '*Ça va pas bien*' ('*I'm not well*')... The entire colloquial corpus is made of 415 sentences: 39 home automation orders, 93 distress sentences, the others are usual sentences. Examples of phrases are given in Table 1.

5. Distress Situation Detection Evaluation

The next sections present the evaluation of the AUDITHIS and RAPHAEL systems. Our evaluation is oriented to distress situation detection. First the results of the evaluation of the sound recognition system and the performances of our ASR in the recording conditions of a flat are assessed. During this experiment, a person is alone in his home and is uttering sentences which are or not distress sentences; this experiment aims to evaluate AUDITHIS and especially the distress keyword detection by RAPHAEL as explained in section 4.6.

Speaker Identifier	1	2	3	4	5	6	7	8	9	10
Sentences with distress keyword (197)	21	19	20	18	20	19	17	21	21	21
Sentences without distress keyword (232)	24	25	23	24	22	24	23	20	24	23

Table 2. Experimental Recorded Corpus: best SNR sentences

Secondly the known source noise suppression system presented in section 4.2 is evaluated in presence of music or radio broadcasting.

5.1 Sound Recognition and Normal/Distress Sentence Recognition

5.1.1 Experiment set up

To validate the system in uncontrolled conditions, we designed a scenario in which every subject had to utter 45 sentences (20 distress sentences, 10 normal sentences and 3 phone conversations made up of 5 sentences each) and to perform different sounds inside the flat. For this experiment, 10 subjects volunteered, 3 women and 7 men (age: 37.2 $\pm$ 14 years, weight: 69 $\pm$ 12 kg, height: 1.72 $\pm$ 0.08 m). To realize these successions of sentences, we chose 30 typical sentences from the colloquial corpus that we randomly scrambled 5 times; then we realized 5 real phone conversations containing 5 successions of sentences, and we picked randomly 3 of the 5 phone conversations.

The experiment took place during daytime – hence we did not control the environmental conditions of the experimental session (such as noises occurring in the hall outside the flat). The sentences were uttered in the flat, with the subject sat down or stood up. The subjects were situated between 1 and 10 meters away from the microphones and have no instruction concerning their orientation with respect to the microphones (they could choose to turn their back to the microphone direction). Microphones were set on the ceiling and directed vertically to the floor. The phone was placed on a table in the living room. The uttered sentences were chosen from the colloquial corpus presented in section 4.6.

Each subject was asked to first enter the flat and close the door, and then to play a scenario (close the toilet door, make a noise with a cup and a spoon, let a box fall on the floor and scream *'Aïe !'*). This whole scenario was repeated 3 times for each subject. Then, he had to go to the living room, and close the communication door between the kitchen and the living room, go to the bed room and read the first half of one of the lists of sentences containing 10 normal and 20 distress sentences. Afterwards, he had to go to the living room and utter the second half of the set of sentences. Each subject was finally called 3 times and must answer the phone and read the phone conversation given (5 sentences each).

Every audio signal was recorded by the application, analyzed on the fly and finally stored on the computer. Each detected signal was first segmented (as sound or speech), and then classified (as one of the eight classes) for a sound, or, in case of a speech event, the 5 more probable hypothesis were stored. For each sound, an XML file was generated, containing the important information.

During this experiment, 2,019 audio signals with an SNR of less than 5 dB were not kept; this 5 dB threshold was chosen because of the poor results given by classification and recognition under this value (Vacher et al., 2006). The number of audio signals collected in this experiment was 3,164 with an SNR of 12.65 $\pm$ 5.6 dB.

After classification, 1008 sounds and 2156 sentences were kept with a mean SNR of 14.4 $\pm$ 6.5 dB. When a sentence was uttered by the speaker, more than one audio signal was recorded

		Results								
		Clap	Step	Phone	Dishes	Lock	Break	Falls	Scream	Speech
Action	Clap	**81.25 %**	0 %	0 %	0 %	0 %	0 %	18.75 %	0 %	0 %
	Phone	0 %	0 %	**100 %**	0 %	0 %	0 %	0 %	0 %	0 %
	Dishes	0 %	0 %	0 %	**42.86 %**	0 %	0 %	0 %	4.76 %	52.38 %
	Falls	19.05 %	0 %	0 %	4.76 %	0 %	0 %	**76.19 %**	0 %	0 %
	Scream	8.7 %	0 %	0 %	8.7 %	0 %	0 %	30.43 %	**30.43 %**	21.74 %

Table 3. Confusion matrix for sound classification (bold values corresponds to the well classified sounds, some sounds are classified as speech).

by the 7 microphones, depending on his position in the room, but only the signal with the best SNR was kept. At the end, the recorded speech corpus was composed of 429 sentences (7.8 minutes of signal), 7 sentences were not kept because of signal saturation (see Table 2). The repartition of the sentences among the speakers is quite well balanced. This corpus was then indexed manually because the speakers did not follow strictly the instructions given at the beginning of the experiment. Moreover, when two sentences were uttered without a sufficient silence between them, some of these couples were considered as one sentence by the audio analysis system. For these reasons, the number of sentences with and without distress keyword was not the same for each speaker.
This first experiment will be used in the two following sections for the evaluation of first the sound recognition and then the identification of distress keywords in the transcribed speech of the AuditHIS system.

5.1.2 Normal/Distress Situation Recognition from Sounds

The results of this experiment are summed-up in the confusion matrix of the global system (Table 3). Rows are the actions performed, and columns give the result. The bold values are the correct decisions that were taken by the system.
This table shows the classes that are (e.g. object fall and doors clapping or dishes and normal sentences) difficult to separate. Some of the classes are very well recognized (phone ringing for instance) but dishes sounds for instance is very difficult to identify, especially because it is not correctly segmented (it is recognized as speech instead of sound). To complete this table, we could add that the global performances of the system are 89.76% of good differentiation sound/speech and 72.14% of well-classified sounds.

5.1.3 Normal/Distress Situation Recognition from Speech

The 429 sentences were analysed by RAPHAEL using both acoustical and language models presented in section 4.6. The uttered sentences were recorded at various input levels depending on the position of the speaker in the room; therefore the dynamic of the signal was increased; the maximal input level for each file was set to 50% of the maximal level if requested. These sentences are distress sentences (DS) or normal sentences (NS).
Distress keywords are extracted by a subsequent process from the complete recognized sentences: it is a Missed Alarm (MA) if the uttered sentence is a distress sentence and if there is no distress keyword in the recognized sentence. In the opposite way, it is a False Alarm (FA) if the uttered sentence is a usual conversation sentence or a home automation order sentence and if the recognized sentence contains a distress keyword. We define the Missed Alarm Rate (MAR), the False Alarm Rate (FAR) and the Global Error Rate (GER) in Equation 3 as:

Error Type	MAR	FAR	GER
Error Rate	29.5%	4%	15.6%

Table 4. Distress Keyword Performance for the Experimental Corpus

$$MAR = \frac{nMA}{nDS} \quad , \quad FAR = \frac{nFA}{nNS} \quad , \quad GER = \frac{nMA + nFA}{nDS + nNS} \tag{3}$$

n referring to the 'number of'.

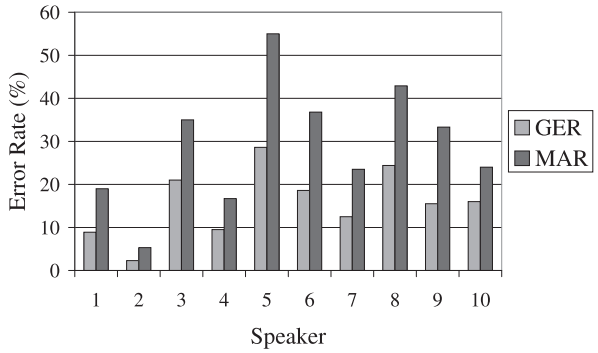

Fig. 5. GER and MAR as a function of the speaker

The results are shown as a function of the speaker on Figure 5 and given overall in Table 4. The FAR is too low to be displayed as a function of the speaker. MAR and GER highly depend on the speaker. For one speaker the MAR is about 5% but for another one, it is above 50%; this speaker uttered distress sentences like a film actor, so some sentences are very different from the sentences of the corpus and this led to errors. For example the French article 'je' was not uttered at the beginning of one sentence. For another speaker, a woman, the MAR is upper 40%. This speaker walked when she uttered the sentences and made noise with her high-heeled shoes, this noise was added to the speech signal. More generally, one distress sentence is *'help!'*, this sentence is well recognized if it was uttered with a French pronunciation but not with an English pronunciation because the phoneme [h] does not exist in French. When a sentence was uttered in the presence of an environmental noise or after a tongue clicking, the first phoneme of the recognized sentence will be preferentially a fricative or an occlusive and the recognition process may be altered.

5.2 Noise Suppression Evaluation

TV and radio signals are perturbing the hearing in the room but a speaker can distinguish easily this noise from speech even if the SNR is about 0 dB. It is not the case for automatic systems, thus it is of great interest to evaluate their performances in conjunction with suppression techniques.

5.2.1 Noise Suppression Experiments

Two microphones were set in a room, the Reference Microphone in front of the Speaker System in order to record music or radio news (radio broadcasting news all the day) and the Signal Microphone in order to record a speaker uttering sentences in the room as shown on Figure 6. The two microphones are connected to AUDITHIS in charge of echo-cancellation; the resulting signal after echo-cancellation with or without post-filtering is then sent to RAPHAEL and stored for further analysis. For this experiment the French speaker is standing in the centre of the recording room, he is not facing the signal microphone. He has to speak with a normal voice level, the level of the radio is set to be rather strong and then the SNR may be approximately 0 dB but there is no real control of the SNR.

Another way is to record separately the reference and the noise after propagation in the room. The speech signal may then be added to the resulting noise at different SNR levels; the Normal/Distress corpus recorded during precedent studies (Vacher et al., 2006) may be used for this purpose. Echo-cancellation is operated in batch accorded to the reference and the addition of speech and noise. So it is possible to proceed with the same signal using different settings of the echo-canceller. This mixing experiment allows a full control of the SNR.

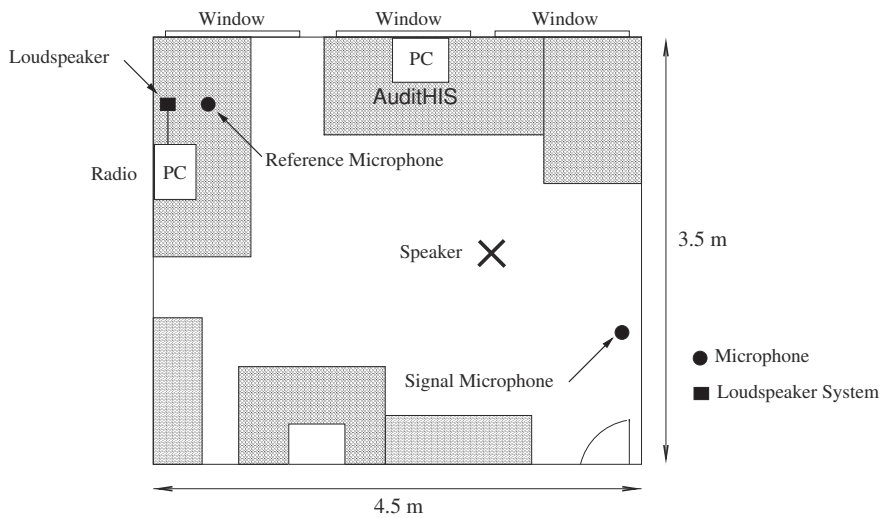

Fig. 6. Setting of the microphones and speaker system in the recording room

5.2.2 Noise Cancellation Results for the Mixing Experiment

The reference signal and the resulting noise in the room were recorded on the 2 microphones during 30 minutes at 16 kHz sampling rate. 126 sentences uttered by one speaker were extracted from the Normal/Distress corpus and mixed with the resulting noise at 9 SNR levels: -12, -9, -6, -3, 0, 3, 6, 9 and 12 dB. Each SNR level was obtained by adjusting the level of both the recorded noise and the audio file of the corpus. The resulting signal was then processed by the echo-cancellation system and the 126 sentences were extracted and sent to the ASR. This process was iterated a second time by the echo-cancellation system with post-filtering. The language model of the ASR was a medium vocabulary statistical system (9958 words in

French). This model is obtained by extraction of textual information from the Internet and from the French newspaper *"Le Monde"*. Then it was optimized using our conversation corpus (refer to Table 1). The recognition results for these two processing methods are presented on Figure 7. The buffer size of the algorithm was 256 samples in order to improve the processing time; the filter size was 8192 samples enough to take into account the size of the room and the delay after reverberation on walls and windows.

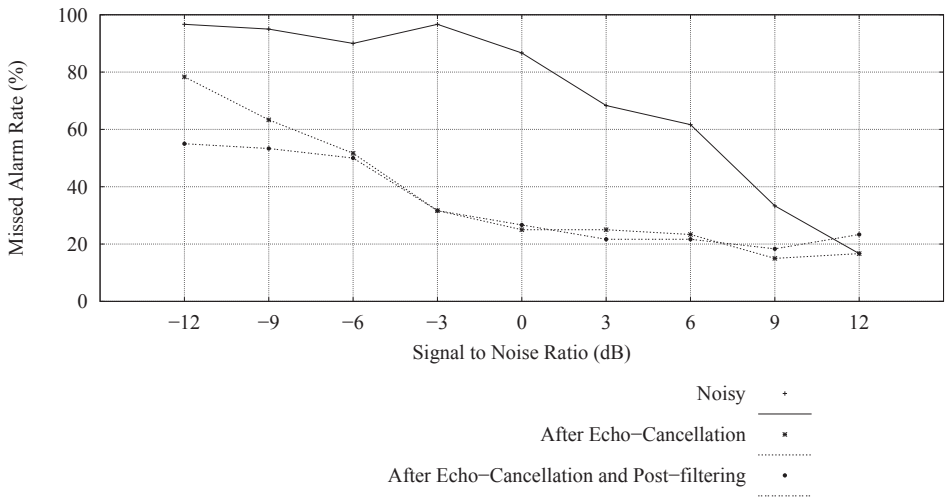

Fig. 7. MAR with France-Info broadcast news

In the absence of echo-cancellation, distress keywords are badly recognized, the MAR is fast increasing under +10 dB SNR. The MAR curve is nearly flat between -3 dB and + 12 dB when echo-cancellation is processed, the post-filtering does not improve significantly speech recognition in this interval, the MAR is even greater at +12 dB. On the contrary, the post-filtering is important below -6 dB and allows the MAR to be 55% (78% for echo-cancellation alone).
The echo-cancellation system was tested with the same corpus with 2 different noise sources: classical music (The 3rd symphony opus 55 by Beethoven) and pop music (Artificial Animals Riding on Neverland by AaRON). Results are displayed on Figure 8, they are very different from the results obtained with the news and the error rate increases linearly with noise level. Thus this kind of noises, and especially pop music, are more difficult to suppress because of the presence of large band sources like percussion instruments.

5.2.3 Noise Cancellation Results for the Real-Time Experiment

In complement, 4 speakers (3 men, 1 woman, between 22 and 55 years old) uttered 20 distress sentences of the Normal/Distress corpus in the recording room, this process was operated by each speaker 2 or 3 times. The echo-cancellation was operated in real-time by the Audio System analysis. Results are shown on Figure 9. The level of the radio France-Info was set in order to achieve a near 0 dB SNR level, each speaker was standing in the centre of the recording room. The MAR, global for all the speakers, is 27%. The results depend on the voice level of the speaker during this experiment and on the speaker himself. Resulting noise at

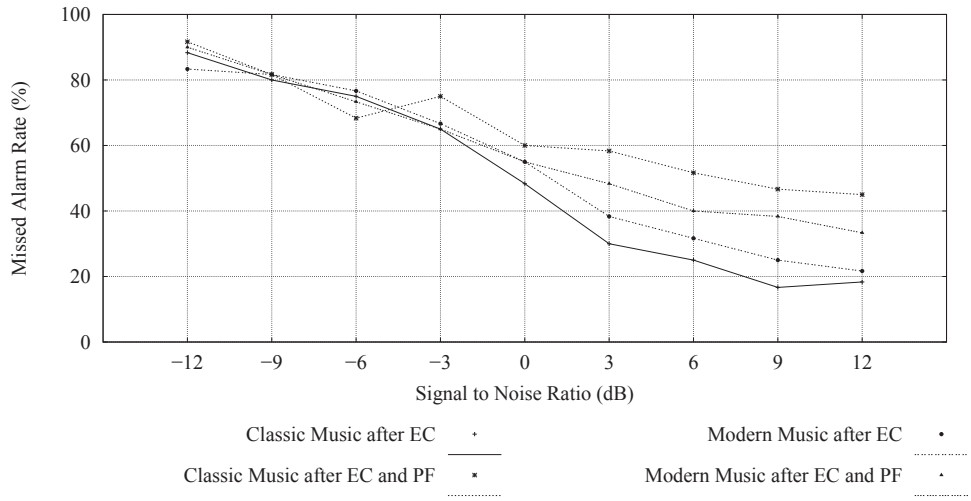

Fig. 8. MAR with Music as Noise Source

the beginning and at the end of the sentence alters the recognition; it may then be useful for detecting these 2 moments with a good precision to use shorter silence intervals.

6. Method for Assessment of the Usefulness of Sound and Speech Recognition for Classification of Activities of Daily Living in Health Smart Homes

To assess the impact of the sound and speech processing on the ADL classification, we have applied several feature selection methods with numerous machine learning schemes. Attribute selection (also called feature selection) is the process of finding a subset of attributes that leads to a good representation of the original data generally for classification (Saeys et al., 2007). Apart from reducing the data set which eases the learning (i.e., less time and memory used) this often leads to better classification performance as useless information (i.e., noise) can degrade the learning phase. This also leads to a more compact classifier representation which is more humanly interpretable (according to the model learned). The usefulness of the audio processing for ADL classification is assessed by the number of audio attributes selected by the attribute selection methods and the impact of this selection on the learning performance. This section presents the machine learning methods retained and the attribute selection techniques applied.

6.1 Supervised Learning

Supervised learning consists in building a classifier *Model* from a learning set L composed of N instances (also called examples or individuals) described by M attributes $A_1 \times A_2 \times \cdots \times A_M$, where A_m represents the domain of the mth attribute, plus a specific attribute C, which represents the class (i.e., concept to learn) in a finite discrete domain (i.e., no regression). The *Model* is then used to predict the class of new unclassified instances (i.e., for which C is unknown). The prediction results, on a specific testing set, is used to assess the *score* of *Model*.

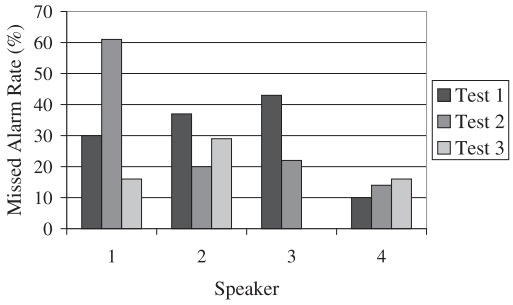

Fig. 9. MAR with Echo-Cancellation in Real-Time

In this work, the learning induction algorithms used are: *Decision Tree (C4.5)*, *Decision Table Majority (DTM)*, *Naïve Bayesian Network (NBayes)* and *Support Vector Machine (SVM)*. They have been chosen for their popularity in data mining applications and because they represent quite different approaches to learning.
Although most of the chosen algorithms can handle numerical attributes (C4.5, NBayes, SVM), L has been discretized using supervised discretization. The method consists in dividing the continuous domain of the attribute with respect to the class into discrete intervals containing the smallest information as possible (i.e., the sub-interval values are similar). As our numerical attributes are derived from categorical ones and as they do not respect normal distribution, this transformation is clearly justified.

6.2 Attribute Selection

Attribute selection techniques are generally divided into two families *filter* and *wrapper* (*embedded* attribute selection is sometimes considered as another family) (Saeys et al., 2007). Broadly speaking, supervised attribute selection by *filter* consist in either ranking each attribute according to its value with respect to C or finding a subset $SA \subseteq A_1 \times A_2 \times \cdots \times A_M$ which describes the most the class C. *Wrapper* method consists in finding a subset $SA \subseteq A_1 \times A_2 \times \cdots \times A_M$ for which the *Model* learned with the target learning algorithm leads to a score better than or close to *Model* learned with L. Single attribute evaluation (for ranking) makes only use of a specific metrics. To test the impact of attribute selection on the learning performance a small subset of each method type has been chosen.

- *Correlation-based Attribute Selection (CorrFA)* searches for subsets of attributes that are highly correlated with the class but with minimal inter-correlation with each other. This method is thus well suited to discover non redundant attributes sets such as one could expect in health smart home.
- *Consistency-based Attribute Selection (ConsFA)* searches for subsets of attributes that are consistent with C. An attribute subset is inconsistent if there are more than one instance with same attribute values but associated with different classes.
- *Wrapping* method is an attribute subset selection method that uses the targeted learning algorithm score at each node as evaluator. This method is more time consuming but leads generally to higher performance than the previous described as it fits the learning algorithm.

Sensors	Attributes	Number	Domain value	Information
Seven Microphones	Nb_sound_w	7	N	Number of times a sound is detected in a room during the time window.
	Nb_sound_x	9	N	Number of times a detected sound has been classified in one of the 9 classes.
Seven Presence Infra-Red sensors	in_y	7	[0,1]	Ratio of time spent in each room (room occupation) during the time window.
	Nb_in_y	7	N	Number of detections in each room (agitation) during the time window
Three Doors contacts	z_state	3	{open, close}	Most used position in the window (Cooking: use of the cupboard and fridge; dressing: use of the dresser).
	Nb_z_open	3	N	Number of times the doors have been opened during the time window.
One in-home weather station	temperature, hygrometry	2	$\mathbf{R}^+$	Differential measure for the last 15 minutes for temperature and hygrometry (use of the shower).

$w \in \{$ kitchen, WC, corridor, bedroom_window, bedroom_wall, living_room_window, living_room_wall $\}$, $x \in \{$ foot_step, dishes_sound, door_closing, door_locker, glass_breaking, speech, scream, phone_ring, object_fall $\}$, $y \in \{$ doorway, bathroom, corridor, WC, living_room, bedroom, kitchen $\}$, $z \in \{$ fridge, dresser, cupboard $\}$.

Table 5. Sensors and their associated attributes and information

7. Evaluation of the Impact of Audio Processing for Activity of Daily Living Classification

The assessment method has been applied to data acquired in the Health Smart Home of Grenoble (see Section 2). The data acquisition and the results of the application of the method on this data are described in the following sections.

7.1 Data Collection in 'Daily Living Conditions'

An experiment has been run to acquire data in the Health smart home of the TIMC-IMAG lab located in the Faculty of Medicine of Grenoble (Fleury, 2008). Thirteen healthy participants (6 women, 7 men) were asked to perform 7 activities, at least once, without condition on the time spent. The average age was 30.4 ± 5.9 years (24-43, min-max), height 1.76 ± 0.08 meters (1.62-1.92, min-max) and weight 69 ± 7.42 kg (57-80, min-max). The mean execution time of the experiment was 51min 40s (23min 11s – 1h 35min 44s, min–max). A visit, previous to the experiment, was organized to ensure that the participants will find all the items necessary to perform the seven ADLs. Participants were free to choose the order with which they wanted to perform the ADLs to avoid repetitive pattern. The 7 activities were defined based on the ADL scale: (1) Sleeping; (2) Resting: watching TV, listening to the radio, reading a magazine. . . ; (3) Dressing and undressing; (4) Feeding: realising and having a meal; (5) Eliminating: going to the toilets; (6) Hygiene activity: washing hands, teeth . . . ; and (7) Communicating: using the phone.

The flat (cf. Fig. 1) contains 18 sensors from which 38 attributes have been derived and are presented in Table 5. Data has been annotated by cutting down each ADL interval into 3-minute windows (the adequate time to perform the shortest activity) labelled with the name of the activity. This process resulted in 232 windows for which: 1) 49 were sleeping; 2) 73

Rank	CorrFA	ConsFA	Global Filter
1	in_bathroom (100%)	in_bathroom (100%)	in_bathroom (100%)
2	in_living_room (100%)	in_WC (100%)	in_living_room (100%)
3	in_bedroom (100%)	in_living_room (100%)	in_bedroom (100%)
4	in_kitchen (100%)	in_bedroom (100%)	in_kitchen (100%)
5	Nb_in_kitchen (100%)	in_kitchen (100%)	Nb_in_kitchen (95%)
6	Nb_sounds_bedroom_-window (90%)	Nb_in_living_room (100%)	in_WC (90%)
7	in_WC (80%)	Nb_in_bedroom (100%)	Nb_sounds_bedroom_-window (85%)
8	Nb_sounds_living_-room_window (80%)	dresser_state (100%)	NB_sound_speech (75%)
9	($\leq$ 50%)	Nb_in_kitchen (100%)	dresser_state (70%)
10		NB_sound_speech (100%)	Nb_sounds_living_-room_window (65%)
11		till the 16th selected	Nb_in_living_room (60%)

Table 6. Attribute ranking for filter attributes selection methods

were resting; 3) 16 were dressing; 4) 45 were eating; 5) 16 were eliminating; 6) 14 were hygiene activity; and 7) 19 were communicating (phone). The final dataset was thus composed of 232 examples of 3-minute activity described by 38 attributes (i.e., attributes) plus the class attribute (i.e., the activity label).

7.2 Attribute Selection Results

Table 6 gives the results of the filter selection method applied on the whole set with a 10-fold stratified cross-validation. For each method the rank of the attribute is given and the number in brackets indicates the percentage of time it has been selected during the cross-validation. The number of retained attributes varies with the method employed but a global trend appears. Four PIR attributes are always in the top of the list (in_bathroom, in_living_room, in_bedroom and in_kitchen) followed by two others (Nb_in_kitchen and in_WC) and by two microphone attributes (sound bedroom, and speech). A rapid analysis of the projection of these attributes against the classes shows that in_bathroom is correlated with hygiene; in_kitchen with eating, in_WC with elimination, in_bedroom with sleeping, dressing and resting, and in_living_room with resting and communicating. This is not surprising as each room is related to several ADLs and thus the presence of someone in a room has a high predictive value about what s/he is doing in it. Regarding the sound attributes, Nb_sounds_bedroom_window and NB_sound_speech are correlated with dressing and communicating. Thus, the audio attributes were found to be informative according to these two activities.

Attribute selection results using the wrapper methods are given in Table 7. The most noticeable changes with the filter method is the dresser_state rank and the disappearance of all bedroom attributes. The former is a Boolean attribute which has a low entropy which is true (i.e., open) only in case of dressing, making it quite interesting for classification. in_bedroom attribute is not discriminative enough to distinguish sleeping from dressing and the configuration of the flat (see Fig. 1) makes this sensor fire sometimes when the participant is resting. Regarding the sound attributes, only Nb_sound_object_fall has been selected as globally useful for ADL classification due to its high correlation with communicating.

Rank	C4.5	DTM	NBayes	SVM (γ=0.8, greedy hill-climbing)	Global Wrapper
1	in_WC (100%)	in_WC (100%)	in_bathroom (100%)	in_bathroom (100%)	in_WC (100%)
2	in_living_-room (100%)	in_living_room (100%)	in_WC (100%)	in_WC (100%)	in_living_room (100%)
3	in_kitchen (100%)	in_kitchen (100%)	in_living_room (100%)	in_living_room (100%)	in_kitchen (100%)
4	dresser_-state (80%)	dresser_state (80%)	in_kitchen (100%)	in_kitchen (100%)	dresser_state (90%)
5	in_bathroom (70%)	Nb_in_bathroom (70%)	dresser_state (100%)	dresser_state (100%)	in_bathroom (77.5%)
6	($\leq$ 40%)	Nb_sounds_liv-ing_room_window (60%)	Nb_sound_object_-fall (80%)	Nb_sound_-object_fall (100%)	Nb_sound_-object_fall (60%)
7		($\leq$ 40%)	Nb_sounds_liv-ing_room_window (80%)	Nb_sound_-door_closing (80%)	
8			Nb_in_bathroom (60%)	($\leq$ 50%)	
9			($\leq$ 50%)		

Table 7. Attribute ranking for wrapper attribute selection methods with best-first search

method	Whole set	No sound	PIR only	Sound only	GF	GW
C4.5	83.28	76.77**	71.65**	51.50**	82.46	83.41
DTM	79.95	75.73**	71.34**	51.20**	82.59	83.02*
NBayes	85.27	77.37**	72.90**	50.78**	85.06	84.49
SVM ($\gamma = 0.8$)	82.91	78.88*	75.00**	51.94**	81.29	84.57
average	82.85	77.19	72.72	51.35	82.85	83.87

$^{*}p < .05$;$^{**}p < .01$

Table 8. Correctly classified Instances (%) for different learning algorithms and data sets

7.3 Impact on the Supervised Learning

The impact of the attribute selection has been assessed by learning from the data sets composed from subsets of attributes using a 10-fold stratified cross-validation repeated 10 times in the same learning conditions as for the wrapping selection. Table 8 summarises the results. Performance with the whole set is the reference for corrected paired student T-test.

Results with the whole set give the lowest performances for DTM (79.95%). NBayes and C4.5 have significantly higher performance than DTM ($p < .05$) but no significant difference is observed between them nor with SVM. When the sound attributes are removed (i.e., 'No sound' data set), the performances are significantly lower. This shows that the sound processing does present essential information to perform activity recognition. Data set composed of PIR attributes (i.e., 'PIR only' data set) only gives significantly reduced performances but still reasonable which emphasizes the location information impact for the classification. Whereas sound sensor should provide information redundant with the PIR sensors at a higher level (foot step, speech), the data set composed only of sound attributes (i.e. 'sound only' data

set) leads to very poor performances. However, the Signal-To-Noise ratio of the sound signal must be improved to reach satisfying performance in this flat which has poor noise insulation (Fleury, 2008). This shows that each modality seems to play a different role for activity classification.

The data set composed of the attributes selected by Global Filtering (GF) attribute selection method leads to performances that are not significantly different from the ones with the whole set. This data set contains less than 29% of the original data (using only 7 sensors). No learning scheme is significantly better than the others. The data set composed of the attributes selected by Global Wrapping (GW) attribute selection method leads to performances that are significantly better for the DTM learning scheme ($p < .05$) but not for the other schemes when compared with the whole set. No significant difference is observed when compared to the GF performances. This data set contains less than 16% of the original data (using only 6 sensors). No learning scheme is significantly better than the others. Overall the GW method leads to higher average performance (83.87%) than with the Whole set (82.85%) and than the GF method (82.85%) but this is not significant and is mainly led by the DTM learning scheme.

The main result of the study is that it is possible to keep high performance for automatic classification of ADL when selecting a relevant subset of attributes. About 33% of the sensors (and less than 16% of the attributes) are enough to classify ADL with same (and sometime superior) performance as with the whole data set. But the retained sensors are of different nature (location, sound, contact door) and thus complement each other. The selected attributes were mainly related to PIR sensors and microphones. While these sensors seem to be the most informative, contact door attached to the dresser was essential for classifying dressing activity. Indeed, the chosen ADLs were all related to a location (e.g.: sleeping in the bedroom, eating in the kitchen ...) and when two activities are usually done in the same room (e.g., sleeping and dressing) a strict location sensor is not enough to distinguish them. Thus, realistic ADLs should include activities in unusual location (e.g., eating while watching TV, sleeping on the sofa ...) to challenge the learning process and acquire more accurate models. This is illustrated by the eliminating and hygiene activities which, due to their natural interrelation (e.g., WC and then washing hand) and the flat configuration, challenged the learning.

Globally, sound sensors attributes have a good predictive power and the study shows that this information is essential for ADL classification. But the study also showed the limit of the current audio processing. Indeed, the sound attribute should deliver the same information as the PIR sensors while adding higher semantic level attributes (speech, footstep ...) but the very hostile sound conditions of the experiment shows that the robustness of the current audio processing needs to be improved. However, the presented results confirmed the information power of this modality at least for support to classical smart home sensors.

8. Discussion

The main outcome of this work is the demonstration that acceptable multisource audio processing performances are reachable, in real time, in a noisy environment. Moreover, although the audio processing performances are still not perfect, the study showed that the audio system (AUDITHIS) provides information essential for automatic identification of activity of daily living. This preliminary work, the first, to the best of our knowledge, that includes real-time sound and speech processing tested in a health smart home, is very encouraging. Of course the approach can be improved in many ways and the experimentations helped us to identify a number of limitations and challenges that need to be overcome.

Regarding the sound processing, the experiments generated very mixed results. Only two classes have been correctly recognized (`Clap` and `Phone`). The `Dishes` is very often confused with `Speech` because the training set doesn't include examples of spoon hitting a cup. This has a fundamental frequency in the spectral band of speech, hence the error of the segmentation module of AUDITHIS. `Scream` is often confused with `Falls of objects` and `Speech`. The former has been learned from a very small training set which explains the misclassification. The latter is more related to a design choice. Indeed, `Scream` is both a sound and a speech and difference between these two is quite vague. For example '*Aïe!*' is an intelligible scream that has been learned by the ASR (RAPHAEL) but a scream could also consist in '*Aaaa!*' which, in this case, should be handled by the sound recognizer. We still not have today a method to handle these situations and investigations about this problem is part of our research agenda. However, most of the poor performances can be explained by the too small training set. Our next goal is to extend this set to include more examples with more variation as well as more classes (such as water sounds and a reject class for unknown or non-frequent sounds). This is mandatory because GMM is a probabilistic approach which relies on a high number of instances to learn correctly each class. With a 'design-for-all' aim — which is reasonable considering that many homes in several countries generate the same kinds of sound — a probabilistic approach seems to be the most suited. However, other approaches, such as (Niessen et al., 2008) which relies on non-statistical techniques are worth being considered and we plan to compare our approach to some of the literature such as the ones base on soundscape (Aucouturier et al., 2007; Guastavino, 2006).

The speech recognition for the detection of distress situation led to interesting results in-lab but still needs to be improved. Indeed, the experiments during which persons were asked to utter normal and distress speech sentences in two different rooms showed how degraded the speech recognition is in hostile condition. The error rate for the recognition of distress sentences in this situation reached 30%. This can be partly explained by the very bad acoustic of the flat (with a mean SNR of about 12 dB instead of 27 dB for the learning corpora). Another limitation of the RAPHAEL system is that its language model is constructed with our colloquial speech corpus. We plan to refine this model by adding expressions and words frequently used in daily conditions. This should give better results and make the distress situation detection tool more adapted to real conditions. The experiments have been performed with young healthy people but it is well known that the voice characteristics vary with the age. Thus, a second challenge is to adapt the acoustic model of RAPHAEL to the target population: the elderly people. We plan to collect speech uttered by a large corpus of elderly people to obtain a system usable for elderly. Other way of improving the speech recognition would be to use acoustic models learned directly from the person which would be more adapted to her voice. However, we believe that a speaker independent approach makes more sense for this kind of application which necessitates to be easily integrable to the flat with a minimal setting period. But, an approach that would automatically adapt the speaker independent models to the person's voice would offer an interesting compromise. The speech processing also offers an interesting approach to deal with the presence of several persons in the flat. This problem is often neglegted, because many works consider that the person lives alone. But an aged person is visited many times and one of the aims of the new technology is to fight loneliness by augmenting the number of daily communications. Thus we need to distinguish what the data that belongs to the monitored person is. Speaker diarization (Fredouille et al., 2006) could be adapted to deal with this problem.

Both sound and speech processings are highly perturbed by audio source of daily living (e.g., TV, radio). The tests of the noise cancellation technique gave a better understanding of the kinds of noise source that may be the most difficult to deal with. The test has been performed only in laboratory and future work includes the test of this method in the health smart home conditions. This will give more conclusive insight about the method performances in this condition.

Sound and speech recognition system is used, *in fine*, for the automatic learning and recognition of activities of daily living together with other modalities. We have shown that the audio modality in health smart home for the recognition of ADL has a significant impact on recognition performance. The evaluation showed that the most informative sensors were the PIR ones. However, the audio channel should provide information not only redundant with the PIR one (e.g., presence in one room) but also higher level (e.g., washing machine spinning). Thus we expect that the improvement of the audio processing will lead to higher performance in activity of daily living recognition. The presented results come from a short number of young and healthy participants in only one flat. Future work should include data from elderly people in more that one flat to test the influence of the disposition of a flat on the results. Such supplementary experimentations would be useful to confirm the relative importance of each sensor. One of the limits of the study is to not have taken the time into account. Indeed, activities such as eating may be repeated at regular time. However, this information is useful only when the resident is following his usual routine and may not challenge the learning to acquire more accurate models. Moreover, as in many data acquired from volunteers outside their home, the time information does not respect the participants' routine. Finally, many applications assess the inhabitant's autonomy by computing the deviance from a routine pattern of activities. But, a deviance may not be a sign of autonomy loss. Microphones can thus play a central role to remove the ambiguity by adding some context detection from the sentences uttered during the day (e.g., multiple voices, plumber intervention . . .).

Audio is not a modality much employed in the domain due to its intrusive nature, but sociological studies (Rialle et al., 2008) have shown that the degree of acceptance of sensors in one's home is idiosyncratic and depend on the level of distress of the person and his relatives. Moreover, the current approach does not record what the speakers are saying but only uses speech processing to detect distress situations. Microphone for sound and speech processing is thus a very interesting sensor to acquire information about human activities which has also the capacity of being used as voice command for domotic purposes.

Finally, it should be emphasized that data acquisition in such environment is a very hard task. Although the number of projects related to ambient assisted living is high, only few datasets are from real data (i.e., aged people with loss of autonomy) and dataset acquisition in this non-standardised domain is a challenge by itself.

9. Conclusion

This chapter presents the AUDITHIS system which performs real-time sound analysis from eight microphone channels in Health Smart Home associated to the autonomous speech analyzer RAPHAEL. The evaluation of AUDITHIS and RAPHAEL in different settings showed that audio modality is very promising to acquire information that are not available through other classical sensors. Audio processing is also the most natural way for a human to interact with his environment. Thus, this approach particularly fits Health Smart Homes that include home automation (e.g., voice command) or other high level interactions (e.g., dialogue). The

originality of the work is also to include sounds of daily living as indicators to distinguish distress from normal situations. First development gave acceptable results for the sound recognition (72% correct classification) and we are working on the reduction of missed-alarm rate to improve performance in the near future.
Although the current system suffers a number of limitations and that we raised numerous challenges that need to be addressed, the pair AUDITHIS and RAPHAEL is, to the best of our knowledge, one of the first serious attempts to build a real-time system that consider sound and speech analysis for ambient assisted living. This work also includes several evaluations on data acquired from volunteers in a real health smart home condition. Further work will include refinement of the acoustic models to adapt the speech recognition to the aged population as well as connexion to home automation systems.

Acknowledgment

The authors would like to thank Hubert Glasson that accomplished an amazing work on AUDITHIS and Noé Guirand who worked on the noise suppression. They also are very grateful to the participants who took part to the different experiments. Thanks are also extended to Christophe Villemazet and the RBI company for their support during the AILISA project.

10. References

Abowd, G., Mynatt, E. & Rodden, T. (2002). The human experience [of ubiquitous computing], *IEEE Pervasive Computing* **1**(1): 48–57.

Adami, A., Hayes, T. & Pavel, M. (2003). Unobtrusive monitoring of sleep patterns, *Proc. 25th Annual Int. Conference of the IEEE-EMBS 2003*, Vol. 2, pp. 1360–1363.

Albinali, F., Davies, N. & Friday, A. (2007). Structural learning of activities from sparse datasets, *5th IEEE Int. Conference on Pervasive Computing and Communications*.

Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B. & Hudson, R. (1999). Recognition of elderly speech and voice-driven document retrieval, *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 145–148.

Aucouturier, J., Defreville, B. & Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music, *Journal of Acoustical Society of America* **122**(2): 881–891.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A. & Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition, *Electronics and Communications in Japan, Part 2, Vol. 87, No. 7, 2004* **87**(2): 49–57.

Bonhomme, S., Campo, E., Estève, D. & Guennec, J. (2008). Prosafe-extended, a telemedicine platform to contribute to medical diagnosis., *J. Telemedicine and Telecare* **14**(3): 116–119.

Chan, M., Estève, D., Escriba, C. & Campo, E. (2008). A review of smart homes- present state and future challenges., *Computer Methods and Programs in Biomedicine* **91**(1): 55–81.

Chen, J., Kam, A. H., Zhang, J., Liu, N. & Shue, L. (2005). Bathroom activity monitoring based on sound, *in* S. B. . Heidelberg (ed.), *Pervasive Computing*, Vol. 3468/2005 of *Lecture Notes in Computer Science*, pp. 47–61.

Clavel, C., Devillers, L., Richard, G., Vasilescu, I. & Ehrette, T. (2007). Detection and analysis of abnormal situations through fear-type acoustic manifestations, *IEEE Trans. on Speech and Audio Processing* **4**: 21–24.

Cohen, I. & Berdugo, B. (2001). Speech enhancement for non-stationary noise environments, *Signal Processing* **81**: 2403–2418.

Cowling, M. (2004). *Non-Speech Environmental Sound Classification System for Autonomous Surveillance*, PhD thesis, Griffith University.

Cowling, M. & Sitte, R. (2003). Comparison of techniques for environmental sound recognition, *Pattern Recognition Letter* **24**(15): 2895–2907.

Dalal, S., Alwan, M., Seifrafi, R., Kell, S. & Brown, D. (2005). A rule-based approach to the analysis of elders' activity data: detection of health and possible emergency conditions, *AAAI 2005 fall symposium, workshop on caring machines: AI in eldercare*.

Davis, K. H., Biddulph, R. & Balashek, S. (1952). Automatic recognition of spoken digits, *The Journal of the Acoustical Society of America* **24**(6): 637–642.
URL: *http://link.aip.org/link/?JAS/24/637/1*

Duchêne, F., Garbay, C. & Rialle, V. (2007). Learning recurrent behaviors from heterogeneous multivariate time-series, *Artificial Intelligence in Medicine* **39**: 25–47.

Duong, T., Phung, D., Bui, H. & Venkatesh, S. (2009). Efficient duration and hierarchical modeling for human activity recognition, *Artificial Intelligence* **173**(7–8): 830–856.

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral emplitude estimator, *IEEE Trans. on Acoustic, speech and Signal Processing* **32**(3): 1109–1121.

Fezari, M. & Bousbia-Salah, M. (2007). Speech and sensor in guiding an electric wheelchair, *Automatic Control and Computer Sciences* **41**(1): 39–43.

Fleury, A. (2008). *Détection de motifs temporels dans les environements multi-perceptifs – Application à la classification des Activités de la Vie Quotidienne d'une Personne Suivie à Domicile par Télémédecine.*, PhD thesis, University Joseph Fourier, Grenoble.

Fredouille, C., Moraru, D., Meignier, S., Bonastre, J.-F. & Besacier, L. (2006). Step-by-step and integrated approaches in broadcast news speaker diarization, *Computer Speech and Language Journal* **20**(2–3): 303–330.

Gauvain, J.-L., Lamel, L.-F. & Eskenazi, M. (1990). Design considerations and text selection for BREF, a large french read-speech corpus, *ICSLP'90*, Kobe, Japan, pp. 1097–1100.

Guastavino, C. (2006). The ideal urban soundscape: investigating the sound quality of french cities, *Acta Acustica* **92**: 945–951.

Gustafsson, S., Martin, R., Jax, P. & Vary, P. (2004). A psychoacoustic approach to combined acoustic echo cancellation and noise reduction, *IEEE Trans. on Speech and Audio Processing* **10**(5): 245–256.

Harma, A., McKinney, M. & Skowronek, J. (2005). Automatic surveillance of the acoustic activity in our living environment, *Proc. IEEE Int. Conference on Multimedia and Expo ICME 2005*, pp. 634–637.

Hong, X., Nugent, C., Mulvenna, M., McClean, S. & Scotney, B. (2008). Evidential fusion of sensor data for activity recognition in smart homes, *Pervasive and Mobile Computing* pp. 1–17.

Ibarz, A., Bauer, G., Casas, R., Marco, A. & Lukowicz, P. (2008). Design and evaluation of a sound based water flow measurement system, *Smart Sensing and Context*, Vol. 5279/2008 of *Lecture Notes in Computer Science*, Springer Verlag, pp. 41–54.

Intille, S. (2002). Designing a home of the future, *IEEE Pervasive Computing* **1**(2): 76–82.

Istrate, D., Castelli, E., Vacher, M., Besacier, L. & Serignat, J.-F. (2006). Information extraction from sound for medical telemonitoring, *IEEE Trans. on Information Technologies in Biomedicine* **10**(2): 264–274.

Istrate, D., Vacher, M. & Serignat, J.-F. (2008). Embedded implementation of distress situation identification through sound analysis, *The Journal on Information Technology in Healthcare* **6**(3): 204–211.

Katz, S. & Akpom, C. (1976). A measure of primary sociobiological functions, *International Journal of Health Services* **6**(3): 493–508.

Kröse, B., van Kasteren, T., Gibson, C. & van den Dool, T. (2008). Care: Context awareness in residences for elderly, *Int. Conference of the Int. Soc. for Gerontechnology*, Pisa, Tuscany, Italy.

Kumiko, O., Mitsuhiro, M., Atsushi, E., Shohei, S. & Reiko, T. (2004). Input support for elderly people using speech recognition, *IEIC Technical Report* **104**(139): 1–6.

LeBellego, G., Noury, N., Virone, G., Mousseau, M. & Demongeot, J. (2006). A model for the measurement of patient activity in a hospital suite, *IEEE Trans. on Information Technology in Biomedicine* **10**(1): 92–99.

Litvak, D., Zigel, Y. & Gannot, I. (2008). Fall detection of elderly through floor vibrations and sound, *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, pp. 4632–4635.

Maunder, D., Ambikairajah, E., Epps, J. & Celler, B. (2008). Dual-microphone sounds of daily life classification for telemonitoring in a noisy environment, *Proc. 30th Annual International Conference of the IEEE-EMBS 2008*, pp. 4636–4639.

Michaut, F. & Bellanger, M. (2005). *Filtrage adaptatif : théorie et algorithmes*, Lavoisier.

Moore, D. & Essa, I. (2002). Recognizing multitasked activities from video using stochastic context-free grammar, *Proc. of American Association of Artificial Intelligence (AAAI) Conference 2002*, Alberta, Canada.

Niessen, M., Van Maanen, L. & Andringa, T. (2008). Disambiguating sounds through context, *Proc. Second IEEE International Conference on Semantic Computing*, pp. 88–95.

Noury, N., Hadidi, T., Laila, M., Fleury, A., Villemazet, C., Rialle, V. & Franco, A. (2008). Level of activity, night and day alternation, and well being measured in a smart hospital suite, *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, pp. 3328–3331.

Noury, N., Villemazet, C., Barralon, P. & Rumeau, P. (2006). Ambient multi-perceptive system for residential health monitoring based on electronic mailings experimentation within the AILISA project, *Proc. 8th Int. Conference on e-Health Networking, Applications and Services HEALTHCOM 2006*, pp. 95–100.

Popescu, M., Li, Y., Skubic, M. & Rantz, M. (2008). An acoustic fall detector system that uses sound height information to reduce the false alarm rate, *Proc. 30th Annual Int. Conference of the IEEE-EMBS 2008*, pp. 4628–4631.

Portet, F., Fleury, A., Vacher, M. & Noury, N. (2009). Determining useful sensors for automatic recognition of activities of daily living in health smart home, *in Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP2009), Verona, Italy* .

Rabiner, L. & Luang, B. (1996). *Digital processing of speech signals*, Prentice-Hall.

Renouard, S., Charbit, M. & Chollet, G. (2003). Vocal interface with a speech memory for dependent people, *Independent Living for Persons with Disabilities* pp. 15–21.

Rialle, V., Ollivet, C., Guigui, C. & Hervé, C. (2008). What do family caregivers of alzheimer's disease patients desire in health smart home technologies? contrasted results of a wide survey, *Methods of Information in Medicine* **47**: 63–69.

Saeys, Y., Inza, I. & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics, *Bioinformatics* **23**: 2507–2517.

Soo, J.-S. & Pang, K. (1990). Multidelay block frequency domain adaptive filter, *IEEE Trans. on Acoustics, Speech and Signal Processing* **38**(2): 373–376.

Takahashi, S.-y., Morimoto, T., Maeda, S. & Tsuruta, N. (2003). Dialogue experiment for elderly people in home health care system, *Text Speech and Dialogue (TSD) 2003*.

Tran, Q. T. & Mynatt, E. D. (2003). What was i cooking? towards déjà vu displays of everyday memory, *Technical report*.

Vacher, M., Fleury, A., Serignat, J.-F., Noury, N. & Glasson, H. (2008). Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment, *The 9thAnnual Conference of the International Speech Communication Association, INTERSPEECH'08 Proceedings*, Brisbane, Australia, pp. 496–499.

Vacher, M., Serignat, J.-F. & Chaillol, S. (2007). Sound classification in a smart room environment: an approach using GMM and HMM methods, *Advances in Spoken Language Technology, SPED 2007 Proceedings*, Iasi, Romania, pp. 135–146.

Vacher, M., Serignat, J.-F., Chaillol, S., Istrate, D. & Popescu, V. (2006). Speech and sound use in a remote monitoring system for health care, *Lecture Notes in Computer Science, Artificial Intelligence, Text Speech and Dialogue, vol. 4188/2006*, Brno, Czech Republic, pp. 711–718.

Valin, J.-M. (2007). On adjusting the learning rate in frequency domain echo cancellation with double talk, *IEEE Trans. on Acoustics, Speech and Signal Processing* **15**(3): 1030–1034.

Valin, J.-M. & Collings, I. B. (2007). A new robust frequency domain echo canceller with closed-loop learning rate adaptation, *IEEE Int. Conference on Acoustics, Speech and Signal Processing, ICASSP'07 Proceedings Vol. 1*, Honolulu, Hawaii, USA, pp. 93–96.

Vaseghi, S. V. (1996). *Advanced Signal Processing and Digital Noise Reduction, 1996*.

Vaufreydaz, D., Bergamini, C., Serignat, J.-F., Besacier, L. & Akbar, M. (2000). A new methodology for speech corpora definition from internet documents, *LREC'2000, 2nd Int. Conference on Language Ressources and Evaluation*, Athens, Greece, pp. 423–426.

Wang, J.-C., Lee, H.-P., Wang, J.-F. & Lin, C.-B. (2008). Robust environmental sound recognition for home automation, *IEEE Trans. on Automation Science and Engineering* **5**(1): 25–31.

Wilpon, J. & Jacobsen, C. (1996). A study of speech recognition for children and the elderly, *IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pp. 349–352.