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] and 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

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.

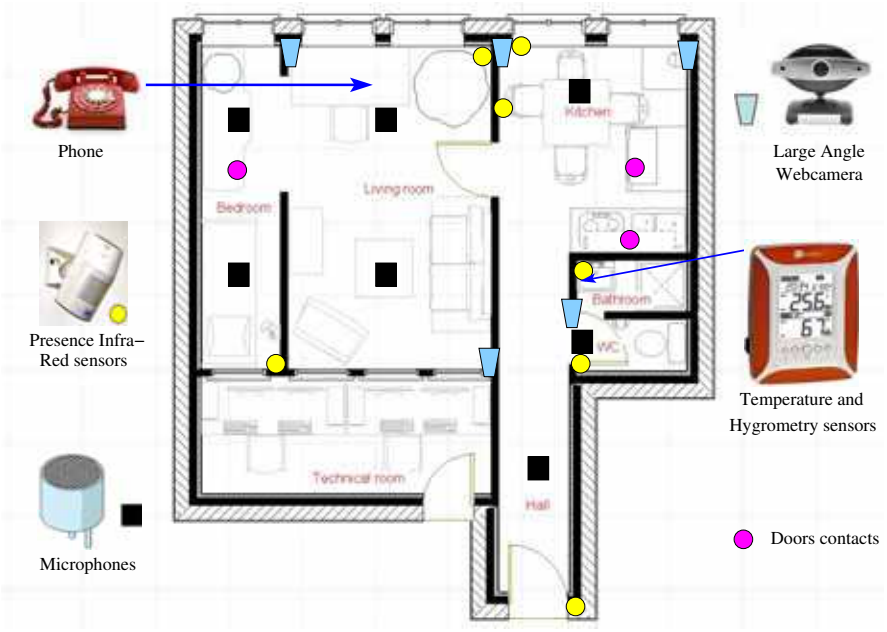

Fig. 1. The Health Smart Home of the TIMC-IMAG Laboratory in Grenoble

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.

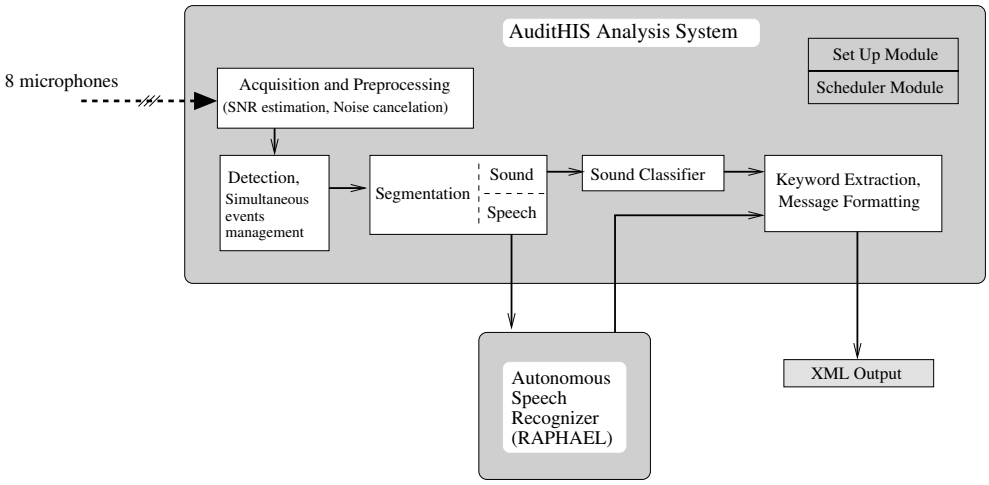

Fig. 2. Architecture of the AUDITHIS and RAPHAEL systems

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.

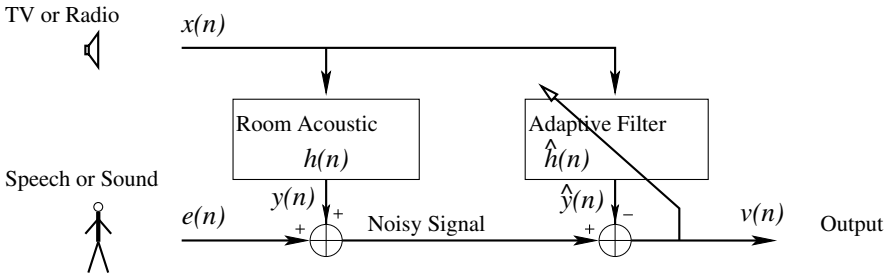

Fig. 3. Echo Cancellation System used for Noise Suppression

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 psycho-acoustical 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]).

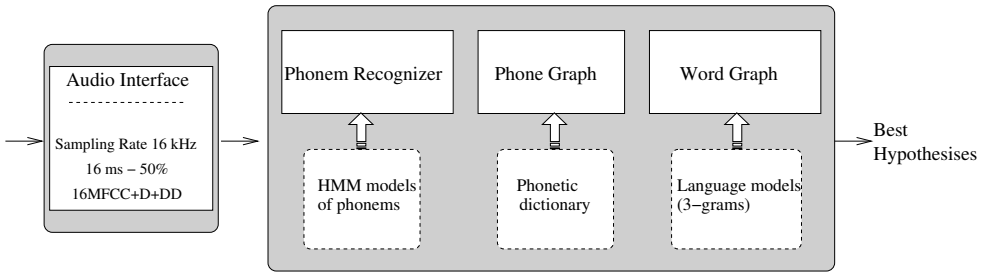

Fig. 4. Architecture of the AUDITHIS and RAPHAEL systems

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.

New Developments in Biomedical Engineering

Edited by Domenico Campolo

ISBN 978-953-7619-57-2
Hard cover, 714 pages
Publisher InTech
Published online 01, January, 2010
Published in print edition January, 2010

Biomedical Engineering is a highly interdisciplinary and well established discipline spanning across engineering, medicine and biology. A single definition of Biomedical Engineering is hardly unanimously accepted but it is often easier to identify what activities are included in it. This volume collects works on recent advances in Biomedical Engineering and provides a bird-view on a very broad field, ranging from purely theoretical frameworks to clinical applications and from diagnosis to treatment.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Michel Vacher, Anthony Fleury, Francois Portet, Jean-Francois Serignat and Norbert Noury (2010). Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living, New Developments in Biomedical Engineering, Domenico Campolo (Ed.), ISBN: 978-953-7619-57-2, InTech, Available from: http://www.intechopen.com/books/new-developments-in-biomedical-engineering/complete-sound-and-speech-recognition-system-for-health-smart-homes-application-to-the-recognition-o

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821