Sound and Speech Detection and Classification in a Health Smart Home

Anthony Fleury, Norbert Noury, Michel Vacher, Hubert Glasson,
Jean-François Serignat

Sound and Speech Detection and Classification in a Health Smart Home

A. Fleury, *Student Member, IEEE*, N. Noury, *Senior Member, IEEE*, M. Vacher, H. Glasson and J.-F. Serignat

***Abstract*—Improvements in medicine increase life expectancy in the world and create a new bottleneck at the entrance of specialized and equipped institutions. To allow elderly people to stay at home, researchers work on ways to monitor them in their own environment, with non-invasive sensors. To meet this goal, smart homes, equipped with lots of sensors, deliver information on the activities of the person and can help detect distress situations. In this paper, we present a global speech and sound recognition system that can be set-up in a flat. We placed eight microphones in the Health Smart Home of Grenoble (a real living flat of $47\mathrm{m}^2$) and we automatically analyze and sort out the different sounds recorded in the flat and the speech uttered (to detect normal or distress french sentences). We introduce the methods for the sound and speech recognition, the post-processing of the data and finally the experimental results obtained in real conditions in the flat.**

***Index Terms*—Sound recognition, Speech recognition, Health Smart Home.**

I. Introduction

DEPENDANCY of elderly persons become an important social problem. Indeed, in France nowadays, 1.3 millions people are over 85 and, in 2015, they will be 2 millions. Loss of autonomy concerns today about 2 millions of people (half are elderly and half are handicapped) and threatens eventually a quarter of the population of elderly people.

For these reasons, geriatrics ask the researchers for tools to automatically detect the decrease in autonomy, so that they can plan the best moment to accept the person in a specialized institution – not to early and not in a hurry. The gain of time before the entrance is one of the solutions to regulate the lack of places in institutions. It is also a chance for the person to live longer in their own environment.

Smart sensors and smart homes have proven their efficiency to give information on the patient. Sensors can deliver Information on postures and movements of the person [1], [2] or detect a fall [3], [4]. Smart Homes are used to measure the activity of the person [5], [6], or to help people (with cognitive impairments for instance) in their activities [7]. Moreover, few projects work on sound and speech recognition systems in smart homes, for automation purpose for instance [8].

This paper describes the implementation and results on the use of microphones for sound classification and speech recognition (in French) in the Health Smart Home of Grenoble. These additional information could be used for two purposes:

- Detect distress situations in the flat by analysing the sounds and also by recognizing a distress sentence,
- By fusing this information with other ones in the flat, we could deduce the activity of daily living actually performed.

II. Materials

A. The Grenoble Health Smart Home

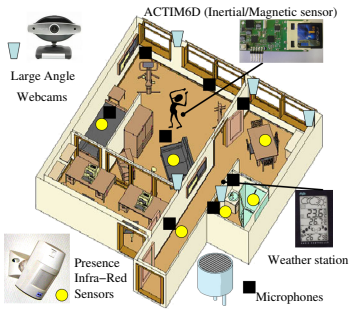

Fig. 1. The Health Smart Home (HIS) at the TIMC-IMAG laboratory, Grenoble

The Health Smart Home set-up by the TIMC-IMAG Laboratory of Grenoble is a real living and equipped flat that measures $47m^2$ (with an equipped kitchen, a bedroom, a living-room, a bathroom) in which the laboratory installed several sensors (see Fig. 1). The sensors used are:

- ▷ Presence infra-red sensors, that give an information on the localization of the patient at a given moment,
- ▷ Open/Close detectors, placed on communication doors and on some other strategic locations such as the door of the fridge or of the cupboard),
- ▷ A weather station delivers hygrometry and temperature,
- ▷ A kinematic sensor, ACTIM6D, placed on the patient, that detects changes in posture (sit-down, lie-down, stand-up, etc.) and provides information on his level of activity,
- ▷ Microphones for sound and speech recognition,
- ▷ Large-angle webcams used only for indexation in learning-based fusion algorithms.

Anthony Fleury and Norbert Noury are with the Laboratory TIMC-IMAG, UMR CNRS/UJF 5525, team AFIRM. Faculté de Médecine de Grenoble, bâtiment Jean Roget, 38706 La Tronche, France (e-mail: {Anthony.Fleury,Norbert.Noury}@imag.fr).

Michel Vacher, Jean-François Serignat and Hubert Glasson are with the LIG Laboratory, UMR UJF/CNRS/INPG 5217, team GETALP. 220 rue de la Chimie, BP 53, 38041 Grenoble Cedex 9. (e-mail: {First-Name.LastName}@imag.fr).

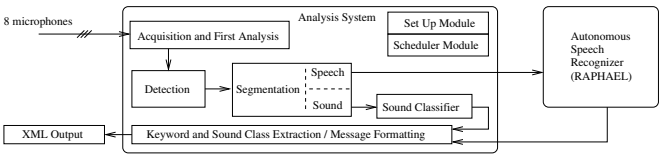

Fig. 2. The global sound and speech recognition system.

All these sensors are linked to the four computers, in the technical room of the HIS, and data are stored on the fly by all of them.

B. Microphones Installation

Eight omnidirectional Electret microphones (ECM-1) have been integrated in the Health Smart Home. They have been placed in the ceiling all around the flat, and have been hidden as much as possible. All of them are plugged to the channel inputs on the acquisition board (NI PCI-6034E, National Instrument), of a computer in the technical room of the HIS.

The microphones are more or less equally distributed in the flat. For instance, in the living-room, that is $2.9m \times 4.75m$, we placed two microphones, at the median axis of the $2.9m$, and about $1m$ from each wall. Two other microphones are placed in the bed-room, one in the kitchen, one in the entrance hall, one in the bathroom and one in the WC.

For each microphones, we have adjusted the gain in the software with the one of the acquisition chain (that depend mostly of the microphone itself) to reach the best dynamic range (maximum detection without saturation).

III. Methods

A. Data Acquisition and Processing

The global organisation of the system is shown on Fig. 2. The following sections introduce the different parts of the system.

1) Sound Detection: The first stage of the sound and speech analysis is the sound detection. The eight analog input channels are continuously and simultaneously sampled by the system at $16kHz$. The noise level is evaluated and the detection of the beginning and the end of the signal use an adaptive threshold [9]. When the beginning and the end of the signal are evaluated, a sound file (wav format) is created ready to be used by the next thread of the application in charge with the segmentation.

2) Sound vs Speech Segmentation: This part has to classify a given sound into speech or sound of daily life. Segmentation is achieved through a GMM classifier trained with the everyday life sound corpus and the normal/distress speech corpus recorded in the LIG laboratory. Acoustical features are LFCC with 16 filter banks and the classifier uses 24 Gaussian models. These features were used because life sounds are better discriminated from speech with constant bandwidth filters than with MFCC and Mel scale. Frame width is 16 ms with an overlap of 50%.

The validation of this segmentation module was made by mixing the sounds and speech records from the corpora and adding them noise recorded in the HIS at 4 Signal to Noise Ratios (training was performed on pure sounds). In these "laboratory" conditions, we obtained a Segmentation Error Rate of 17.3% for a SNR of $0\ dB$, 5.1% at $10\ dB$, 3.8% at $20\ dB$ and finally 3.6% at $40\ dB$. We can notice that SER remains quite constant with a 5% value above $10\ dB$.

3) Sound Classification: When segmented as sounds, the wav file is then processed by the classification part of the algorithm. Everyday life sounds are classified with a GMM or HMM classifier, the classifier is selected before the beginning of the experiment. They were trained with the eight classes of the everyday life sound corpus using LFCC features (24 filter banks) and 12 Gaussian models. The training step is more described in [10] for the GMM method (Expectation Maximisation algorithm) and for the HMM method (algorith of Viterbi).

The every day life sounds are divided into 8 classes corresponding to 2 categories: *normal* sounds related to usual activities of the patient (door clapping, phone ringing, step sounds, dishes sounds, door lock), *abnormal* sounds related to distress situations (breaking glass, fall of an object, screams). This corpus contains some records made at LIG laboratory (61%) using super-cardioids microphones (eW500, Sennheiser), some files coming from a preceding corpus recorded at the time of former studies in the CLIPS laboratory and some files obtained from the Web. The corpus is constituted of 1,985 audio files for a total duration of 35 min 38 s, each file contains one sound.

We also evaluated the performance of this classification, in the same conditions as for segmentation, using different SNR. With the GMM model, 24 LFCC, the Classification Error Rate is 36.6% at $0\ dB$, 21.3% at $10\ dB$, 12% at $20\ dB$ and finally 9.3% at $40\ dB$. We notice again that the error is highly dependant of the SNR.

4) Speech Recognition: The autonomous speech recognizer RAPHAEL [11] is running as an independent application and analyzes the speech events resulting from the segmentation module through a file exchange protocol. As soon as the requested file has been analyzed, it is deleted and the 5 best hypothesis are stored in a hypothesis file. This event allows the scheduler to send another file to be analyzed. The language model of this system is a medium vocabulary statistical system (around 11,000 words in French). This model was obtained by extraction of textual information from the Internet and from the French journal "Le Monde" corpora.

In order to train and validate the system we have recorded two adapted corpora: the *normal/distress speech corpus* in French and the *life sound corpus*. For speech recognition, the training of the acoustic models was made with large corpora in order to insure a good speaker independence. They were recorded by 300 French speakers in the CLIPS laboratory (BRAF100) and LIMSI laboratory (BREF80 and BREF120). All corpora were recorded using the same 16 kHz sampling rate as the analysis system. We have 66 normal and

60 distress sentences [10].

The speech recognition system was evaluated with the sentences from 5 speakers of our corpus (630 tests). In 6% of the cases, for normal sentences, an unexpected distress keyword is detected by the system and leads to a *False Alarm Sentence*. In 16% of the cases, for distress sentences, the distress keyword is not recognized (missed): this leads to a *Missed Alarm Sentence*. This often occurs in isolated words like "Aïe" (Ouch) or "SOS" or in syntactically incorrect French expressions like "Ça va pas bien" (I don't feel well). The global Speech Recognition Error Rate is then 11%.

B. Acquisition Software

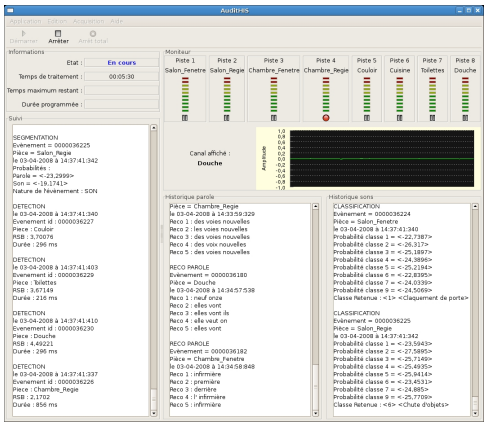

Fig. 3. The interface of the Acquisition software developed at the LIG Laboratory with: 3 text columns containing from left to right (1) the information on the detections, (2) the speech recognition and (3) the sound classification. The level of the eight microphones is represented on the right and the wave of the last detected microphone is continuously drawn.

Fig. 3 presents the application realised under GNU/Linux that implements all the preceding algorithms. This application is a multi-threading application that performs the following tasks: communication with the NI Board, detection of the sounds and creation the Wav files, then segmentation and classification of a sound, or for a speech, communication with the speech recognition system (that is an independent application). For each sounds and speech detected and classified by the software, an XML file containing all the information (date, time, SNR, segmentation and classification or sentences) is created with the associated logs and wav files.

This application allows us both to realize real experimentations and to post-process the data using the created XML files. Moreover, in addition to all that has been described, it has a modifiable threshold so that every sounds under a SNR will be ignored. This allows us to reduce the amount of data to be processed by ignoring the sounds under 5 dB that would be undoubtly very badly classified. These files are not segmented.

C. Post-processing

Each sound is recorded by the application and stored on the hard drive of a computer, with the associated XML file containing the information on the file (from detection to classification). Afterwards, these collected data are processed using Matlab™.

Then the sounds are classified considering and fusing the results of the different microphones using the following algorithm. For a sound that will be done in the flat, we will take the SNR of the best microphone (named x), and keep all the microphones having a SNR greater that $0.8 * x$. We further take the decision from a vote between these different decisions. We apply two rules in case of equality: (1) if a distress speech is detected, we keep this decision and (2) in case of equality with another decision than a distress speech, we keep the decision of the microphone having the best SNR. This classification will give us two pieces of information for each event: the kind of event (sound or speech) and the retained class. We will create a succession of sound and speech events for future use in data fusion.

IV. Experimental Results

A. Protocol

To validate the system in unsupervised conditions, we built a scenario in which every subject has to pronounce 45 sentences (20 distress, 10 normal and 3 phone conversations of 5 sentences each). For this experimentation, 13 subjects volunteered, 3 women and 10 men (age: 33 ± 12 $years$, weight: 64 ± 20 kgs, height: 1.74 ± 0.06 m). The number of sounds collected by this experimentation was $5,417$ ($2,399$ of them were not segmented because their SNR was less that 5 dB), with an SNR of 12.5 ± 5.6 dB. After classification, we kept $1,820$ sounds with a mean SNR of 13.6 ± 6.5 dB.

The experimentation took place during daytime – so we do not control the environmental conditions of the experimental session (such as all the noises in the neighbourhood). The sentences were uttered in the flat, with the subject sitting or standing. He was between 1 and 10 meters away from the microphones and had no instructions on his orientation with respect to the microphones (he could choose to turn his back to the microphone).

The protocol was quite simple. The subject was asked to first make a little scenario (close a door, make a noise with a cup and a spoon, let a box fall on the floor and scream "Aie"). This whole scenario was repeated 3 times. Then, he had to read a succession of 10 normal and 20 distress sentences. After, he received a phone call and had to answer and read the given phone conversation. To realise the five different successions of sentences, we choose 30 representative ones and realised 5 phone conversations, and then we scrambled the sentences five times, and we randomly chose 3 of the 5 conversations for a given subject. This leads to a large number of 563 sentences uttered, out of which 268 are distress sentences.

B. Results

The results of this experimentation are summed-up in the confusion matrix of the global system (Table I). The

TABLE I
CONFUSION MATRIX FOR SOUND AND SPEECH RECOGNITION (BOLD VALUES CORRESPONDS TO THE WELL CLASSIFIED SOUNDS).

		Results									
		Clap	Step	Phone	Dishes	Lock	Break	Falls	Scream	Normal Speech	Distress Speech
Action	Doors Clapping	**81.25 %**	0 %	0 %	0 %	0 %	0 %	18.75 %	0 %	0 %	0 %
	Phone Ringing	0 %	0 %	**100 %**	0 %	0 %	0 %	0 %	0 %	0 %	0 %
	Dishes Sound	0 %	0 %	0 %	**42.86 %**	0 %	0 %	0 %	4.76 %	52.38 %	0 %
	Object Fall	19.05 %	0 %	0 %	4.76 %	0 %	0 %	**76.19 %**	0 %	0 %	0 %
	Scream	8.7 %	0 %	0 %	8.7 %	0 %	0 %	30.43 %	**30.43 %**	21.74 %	0.00 %
	Normal Speech	0.74 %	0 %	0.37 %	4.1 %	0 %	0 %	3.35 %	4.48 %	**83.44 %**	3.49 %
	Distress Speech	0.74 %	0.37 %	0 %	2.4 %	0.37 %	0 %	3.35 %	0 %	62.92 %	**29.85 %**

different lines are the action performed, and the columns give the result of the system. The bold values are the correct decisions that were taken by the system. The action part of the confusion matrix is not complete. As far as the "break" class is concerned, it was difficult to realize such an action during an experiment with 13 subjects and a sufficient number of realization. Additionally, the shoes worn by the subjects did not produce sufficient signal level to be detected.

This table shows us the classes that are close (e.g. object fall and doors clapping or dishes and normal sentences) and difficult to separate. We note 0% between screams and distress sentences, due to the fact that distress sentences could be reduced to a short word uttered by the subject like a scream. It is neither a bad segmentation nor a bad classification to take a scream instead of a distress sentence in this case. Screams are also close to object falls and speech (normal sentences). To complete this table, we could add that the global performances of the system are 89.76% of good segmentation, 72.14% of well-classified sounds and 41% of well-recognized sentences. This leads to 18.1% of false alarms and unfortunately to 70.1% of missed alarms. For the detection part, with our adaptive threshold, each sound is well detected by the system.

V. DISCUSSION AND CONCLUSION

This paper presents a complete sound and speech recognition system, with evaluation results in unsupervised and real conditions, compared to the results obtained in laboratory conditions. For the events tested, we can see that the results for the sound recognition are good and conform with the results obtained in laboratory conditions, if we consider the SNR of the HIS.

As far as speech recognition is concerned, the results are too low, especially for the distress sentence recognition. Even if the corpora was made independent of the speaker, we face difficulties of recognition because each subject pronounces differently the sentence. Moreover, the acquisition line, the microphones and the environment are all imperfect. We could have a noise added to the sound and disturb the HMM process. The conditions are also uncontrolled because the subject could pronounce the sentence when he decided, and could freely choose his orientation to the microphone. Thus our conditions are the worst possible, far from the laboratory conditions (no noise and the microphone just behind the subject). Neverthless these real collected sounds will be used to improve the models of language and the results for next experimentations. We are also working on the learning of other classes.

ACKNOWLEDGMENT

The authors would like to thanks all the subject from both laboratories for their time spent doing the experiments.

REFERENCES

[1] P. Barralon, N. Noury, and N. Vuillerme, "Classification of daily physical activities from a single kinematic sensor," in *27th Annual International Conference of the IEEE-EMBS 2005*, 2005, pp. 2447–2450.

[2] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Büla, and P. Robert, "Ambulatory system for human motion analysis using a kinematic sensor: Monitoring of daily physical activity in the elderly," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 6, pp. 711 – 723, June 2003.

[3] N. Noury, A. Fleury, P. Rumeau, A. Bourke, G. O. Laighin, V. Rialle, and J. Lundy, "Fall detection - principles and methods," in *29th Annual International Conference of the IEEE-EMBS 2007.*, 22-26 Aug. 2007, pp. 1663–1666.

[4] A. K. Bourke, J. V. O'brien, and G. M. Lyons, "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm." *Gait Posture*, vol. 26, no. 2, pp. 194–199, Jul 2007.

[5] G. LeBellego, N. Noury, G. Virone, M. Mousseau, and J. Demongeot, "A model for the measurement of patient activity in a hospital suite," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 92–99, Jan. 2006.

[6] V. Ricquebourg, D. Menga, D. Durand, B. Marhic, L. Delahoche, and C. Loge, "The smart home concept : our immediate future," in *E-Learning in Industrial Electronics, 2006 1ST IEEE International Conference on*, 18-20 Dec. 2006, pp. 23–28.

[7] B. Bouchard, A. Bouzouane, and S. Giroux, "A smart home agent for plan recognition of cognitivly-impaired patients," *Journal of Computers*, vol. 1, no. 5, pp. 53 – 62, Aug. 2006.

[8] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 25–31, Jan. 2008.

[9] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat, "Information extraction from sound for medical telemonitoring," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 10, no. 2, pp. 264 – 274, Apr. 2006.

[10] M. Vacher, J.-F. Serignat, S. Chaillol, D. Istrate, and V. Popescu, "Speech and sound use in a remote monitoring system for health care," in *LNAI, Text, Speech and Dialogue*, vol. 4188, 2006, pp. 711 – 718.

[11] M. Akbar and J. Caelen, "Parole et traduction automatique : le module de reconnaissance raphael," in *Proc. COLING-ACL'98, Montréal, Quebec*, Aug. 10–14 1998, pp. 36 – 40.