Speech recognition in a smart home: some experiments for telemonitoring

Michel Vacher, Noe Guirand, Jean-François Serignat, Anthony Fleury, Norbert Noury

Speech Recognition in a Smart Home: Some Experiments for Telemonitoring

Michel Vacher, Noé Guirand, Jean-François Serignat
LIG Laboratory, GETALP Team
UMR CNRS-UJF-INPG-UPMF 5217
BP53, 38041 Grenoble cedex9, France
Michel.Vacher@imag.fr

Anthony Fleury, Norbert Noury
TIMC-IMAG Laboratory, AFIRM Team
UMR CNRS-UJF 5525
Domaine de la Merci, 38710 La Tronche, France

***Abstract*— Because of the aging of the population, low-cost solutions are required to help people with loss of autonomy staying at home rather than in public health centers. One solution is to assist human operators with smart information systems. In this case, position and physiologic sensors already give important information, but there are few studies about the utility of sound in patient's habitation. However, sound classification and speech recognition may greatly increase the versatility of such a system: this will be provided by detecting short sentences or words that could characterize a distress situation for the patient. Moreover, analysis and classification of sounds emitted in patient's habitation may be useful for patient's activity monitoring. In this paper, we present a global speech and sound recognition system that can be set-up in a flat. Eight microphones were placed in the Health Smart Home of Grenoble (named HIS, a real living flat of 47m2) to automatically analyze and classify different sounds and speech utterances (e.g.: normal or distress French sentences). Sounds are clustered in eight classes but this aspect is not discussed in this paper. For speech signals, an input utterance is recognized and a subsequent process classifies it in normal or distress, by analysing the presence of distress keywords. An experimental protocol was defined and then this system has been evaluated in uncontrolled conditions in which heterogeneous speakers were asked to utter predetermined sentences in the HIS. The results of this experiment, where ten subjects were involved, are presented. The Global Error Rate was 15.6%. Moreover, noise suppression techniques were incorporated in the speech and sound recognition system in order to suppress the noise emitted by known sources like TV or radio. An experimental protocol was defined and tested by four speakers in real conditions inside a room. Finally, we discuss the results of this experiment as a function of the noise source: speech or music.**

Noise Suppression; Smart Home, Speech Recognition; Telemonitoring (key words)

I. INTRODUCTION

The constant growing of life expectancy in the world yields a lack of places and workers in institutions able to take care of elderly people. Researcher teams all over the world try to tackle this issue by working on ways to maintain elderly people in their own home as long as possible. Geriatrics is thus in great need for systems assessing the evolution of a person in her environment and detecting the appropriate moment for admitting that person in an institution. Smart homes can be classified according to the types of equipment and systems installed. The residents of smart homes are not just those with severe pathologies or chronic illness, but also those who want a better quality of life [1]. The most important targets are to maintain a certain level of independence, to continuously monitor vital parameters, to reduce accidents and to help out with delivering therapy.

Abnormal situations in the behaviour of the person should be detected through the information delivered by smart sensors [2]. Smart homes have demonstrated that measuring the activity of a person at home can be relevant [3], and also that this monitoring is useful for people with cognitive impairments [4]. A few studies are related to sound recognition capabilities [5], [6].

In our study, a fully functional flat has been fitted with numerous sensors, chosen for classifying the different activities of a person's everyday life. This 47 m^s flat is located at the Faculty of Medicine of Grenoble. It includes a kitchen, toilet, a bedroom, a living room and a bath room. It is fitted with: -Presence Infrared sensors (PIR) to approximately determine the location of the subject, -a weather station that provides information on temperature and hygrometry, -open/close detectors placed on communication doors, fridge... -an embedded kinematic sensor, -and, finally, eight microphones (one or two per room) who are in the focus of this paper. Large angle webcams have also been placed but are only used to time-stamp the activities realized in the flat (for machine learning applications) and not as sensors. The microphone setting in the flat is shown in Figure 1; microphones are set on the ceiling and directed vertically to the floor. The sentences uttered by the subject, as well as emitted life sounds, may give valuable information on her usual activities, or on a distress situation.

Data from all these sensors are acquired and processed in real time on four computers disposed in the technical room; they are used as inputs to off-line data fusion algorithms, for detecting and classifying daily activities. An audio analysis system is running on a computer and is delivering information in real time; the detected sound and speech signals are stored for further analysis. With regard to speech signals, an input utterance is recognized and a subsequent process classifies it as normal or distress, by analysing the presence of distress key

words. This audio analysis system is presented in section 2, as well as the Autonomous Speech Recognizer (ASR) in section 3; the experiment made in the flat, in order to assess its performances out of "laboratory conditions" is presented in section 4 and the corresponding normal/distress sentence recognition results are given in section 5. Noise suppression techniques in the case of known sources are presented in section 6 and our experiments in section 7; the experimental results are discussed in section 8 before the general conclusion of this study in section 9.

II. THE AUDIO ANALYSIS SYSTEM

Figure 2 depicts the general organization of the audio analysis system; a more detailed description of the system is given in [7]. Each microphone is connected to an analog input channel of the acquisition board; each sound is processed independently thanks to a queuing management protocol. The analysis system and the autonomous speech recognizer are running in real time as independent applications on the same GNU/Linux computer. These two applications are synchronized through a file exchange protocol. The system is made of several modules: -acquisition and first analysis, -detection, -segmentation, -classification, -and finally, message formatting. They run as independent threads synchronized by a scheduler.

Data acquisition is operated on the 8 input channels simultaneously at a 16 kHz sampling rate by the acquisition and first analysis module. Noise level is 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. The detection module is in charge of signal extraction, i.e. to detect the beginning and the end of the audio event.

The segmentation module is a Gaussian Mixture Model (GMM) classifier that classifies each audio event as everyday life sound or speech. The segmentation module was trained with an everyday life sound corpus [8] and with the Normal/Distress speech corpus recorded in our laboratory [7]. 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 [8]. Frame width is of 16 ms, with an overlap of 50%.

Then, the signal is transferred by the segmentation module to the speech recognition system in case of speech or to the sound classifier in case of everyday life sounds. Everyday life sounds are classified with a GMM or Hidden Markov Model (HMM) classifier; the classifier is chosen at the beginning of the experiment. Their models were trained with our corpus containing the eight classes of everyday life sounds, using LFCC features (24 filter banks) and 12 Gaussian models.

III. THE AUTONOMOUS SPEECH RECOGNIZER

The autonomous speech recognizer RAPHAEL [7] 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 hypotheses are stored in a file. This event allows the scheduler to send the next queued file to the recognizer. Moreover, each sentence file is stored in order to allow for future analysis with different recognition parameters for the recognizer.

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 in the CLIPS (BRAF100) [19] and LIMSI laboratories (BREF80 and BREF120) [9].

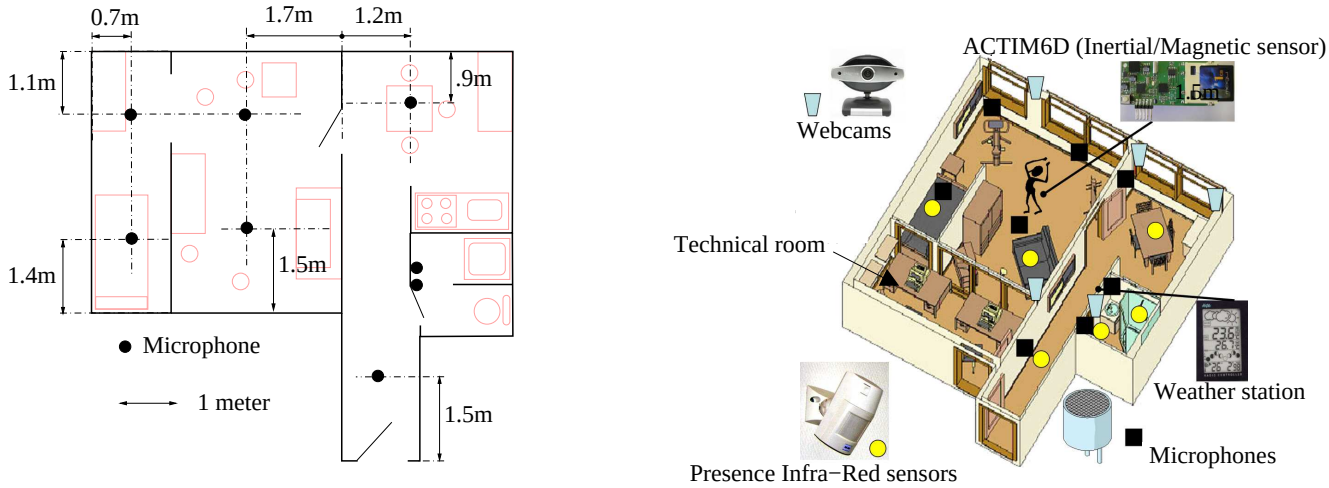


Figure 1. Microphone setting in the flat on the left and full sensors set-up on the right

The language model of this system is a small vocabulary statistical system (299 words in French). This model is obtained using textual information of a current conversation corpus in French. Our main requirement is the correct detection of a possible distress, situation through keyword detection, without understanding the patient's conversation. This conversation corpus contains the sentences in the Normal/Distress speech corpus [10], along with sentences currently uttered during a phone conversation: "*Allo oui*", "*A demain*", "*J'ai bu ma tisane*", "*Au revoir*"... and sentences that may be a command for a domotic system. The Normal/Distress speech corpus is composed of 126 sentences in French: 66 are typical for a normal situation for the patient: "*Bonjour*" (Hello), "*Où est le sel*" (Where is the salt)... , 60 are typical for a distress situation: "*Aouh*", "*Aïe*", "*Au secours*" (Help), "*Un médecin vite*" (Call a doctor hurry) along with syntactically incorrect French expressions like "*Ça va pas bien*" (I don't feel good)... The entire conversation corpus is made of 39 domotic orders, 93 distress sentences and usual conversation sentences. Ten samples of each kind are given in Table 1.

TABLE 1. THE CURRENT CONVERSATION CORPUS

Sample	**Domotic Order Sentence**	**Distress Sentence**	**Usual Conversation Sentence**
1	*Allume la lumière*	*A l'aide*	*Allô c'est moi*
2	*Éteins la lumière*	*Je suis tombé*	*Allô 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*

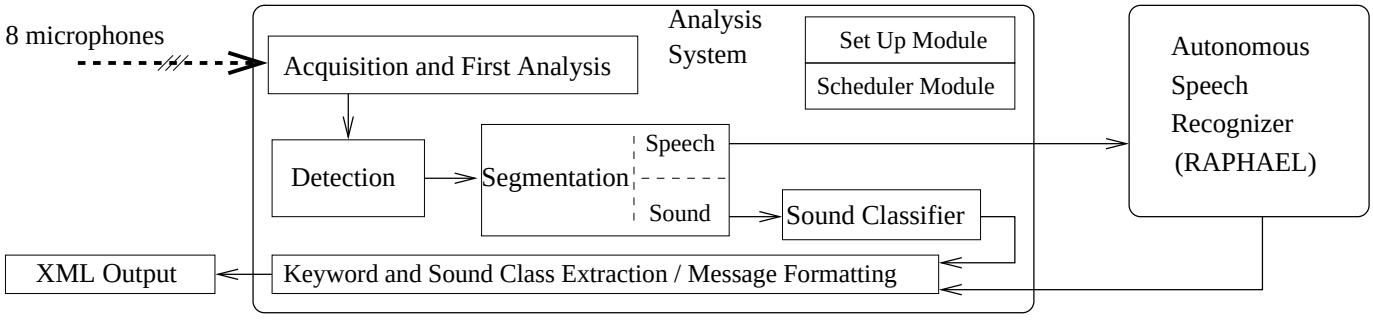


Figure 2. Overview of the Audio Analysis System

IV. EXPERIMENTS AND RECORDED CORPUS

To validate the system in uncontrolled conditions, we designed a scenario where every subject had to utter 45 sentences (20 distress sentences, 10 normal sentences and 3 phone conversations made up of 5 sentences each). Ten subjects -3 women and 7 men (age: 37.2 ± 14 years, weight: 69 ± 12 kg, height: 1.72 ± 0.08 m)- volunteer for this experiment.

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 had no instruction concerning their orientation with respect to the microphones (they could choose to turn their back to the microphone direction). Microphones are set on the ceiling and directed vertically to the floor as shown on Figure 1. The phone was placed on a table in the living room.

The protocol was quite simple. The subjects were asked to first enter the flat and close the door, and then to act a little 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, the subjects had first to go to the living room and close the communication door (between the kitchen and the living room) and then to go to the bed room and read the first half of one of the successions of sentences containing 10 normal and 20 distress sentences. Afterwards, they 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 had to answer the phone and read the phone conversation given (5 sentences each). To realize these successions of sentences, we chose 30 typical sentences 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.

Every audio signal was recorded by the application, analyzed on the fly and finally stored on the hard disk drive of the computer. For each detected signal, it 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 hypotheses were stored. For each sound, a XML file was generated, containing the important information.

During this experiment, 2,019 audio signals with an SNR 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 [10]. The number of audio signals collected in this experiment was 3,164 with an SNR of 12.65 ± 5.6 dB.

After classification, we kept 1,008 sounds with a mean SNR of 14.4 ± 6.5 dB and 2,156 sentences. As part of the study presented in this paper, only the recorded sentences are considered. Recorded sounds such as steps, doors clapping were not taken into account in this study. When a sentence was uttered by the speaker, more than one audio signal was recorded by the 7 microphones, depending on his position in the room, but we chose to keep only the signal with the best SNR. 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). This corpus was indexed manually because each speaker doesn't 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.

TABLE 2. EXPERIMENTAL RECORDED CORPUS: BEST SNR SENTENCES

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

V. NORMAL/DISTRESS SENTENCE RECOGNITION

The 429 sentences were analyzed by the RAPHAEL speech recognizer using the acoustical models and the language model presented in Section 3. These 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 to 50% of the maximal input level for each file below 50% of the maximal level. The language model is made of 299 unigrams, 729 bigrams and 862 trigrams. Afterwards, 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 domotic 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 (1) and (2), *n* referring to the '*number of*', *DS* to '*Distress Sentence*', *NS* to '*Normal Sentence*'.

$$MAR = \frac{nMA}{nDS} \quad , \qquad FAR = \frac{nFA}{nNS} \tag{1}$$

$$GER = \frac{nMA + nFA}{nDS + nNS} \tag{2}$$

The results are shown as a function of the speaker on Figure 3 and given overall in Table 3. As far as the FAR is concerned, the error is low and then not displayed as a function of the speaker.

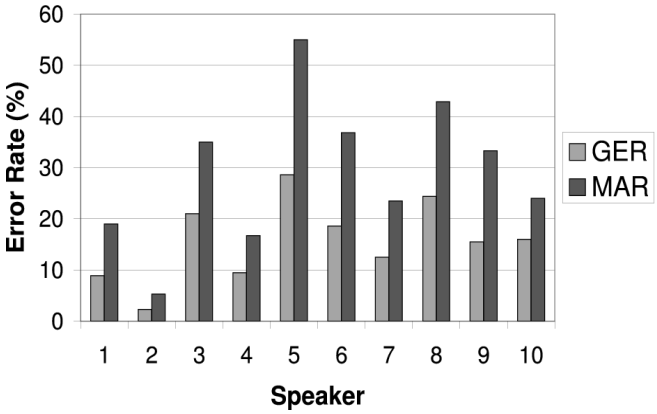


Figure 3. GER and MAR as a function of the speaker

The results, MAR and GER, are very dependent 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, therefore some sentences are very different from the sentences of the corpus and this leads to an error. For example the French pronoun "*je*" was not uttered at the beginning of one sentence. For another speaker, a woman, the MAR is upper than 40%. This speaker walked when she uttered the sentences and made noise with their 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] doesn't 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.

VI. NOISE SUPPRESSION IN THE CASE OF KNOWN SOURCES

Sound emitted by a radio or a TV in the HIS $x(n)$ is a noise source that will be altered by the room acoustic through his transfer function: $y(n) = h(n) * x(n)$, where n represents the discrete time, $h(n)$ the impulse response of the room acoustic and * the convolution operator. It is then superposed to the signal $e(n)$ emitted in the room: speech uttered by the patient or everyday life sound. The signal recorded by the microphone in the HIS is then $y(n) = e(n) + h(n) * x(n)$ *y(n).* Various methods were developed in order to suppress the noise [11], some methods to obtain an estimation $\hat{h}(n)$ of the impulse response of the room acoustic in order to remove the noise as shown on Figure 4.

The resulting output is given in (3).

$$v(n) = e(n) + y(n) - \hat{y}(n)$$

$$v(n) = e(n) + h(n) * x(n) - \hat{h}(n) * x(n) \tag{3}$$

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 [11]. The Multi-delay Block Frequency Domain (MDF) algorithm is an implementation of the LMS algorithm in the frequency domain [12]. This algorithm is implemented in the SPEEX library under GPL License [13] for echo cancellation system.

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 [14] 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.

TABLE 3. DISTRESS KEYWORD PERFORMANCE FOR THE EXPERIMENTAL CORPUS

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

The echo cancellation technique used introduces a specific noise into the *v(n)* signal and a post-filtering is requested. The method implemented in SPEEX is Minimum Mean Square Estimator Short-Time Amplitude Spectrum Estimator (MMSE-STSA) presented in [15]. 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 independent, - the phase of the Discrete Fourier Transform follows a uniform distribution law and is amplitude independent. Some improvements are added to the SNR estimation [16] and a psycho-acoustical approach for post-filtering [17]; 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 [18]. The post-filter is implemented in the frequency domain, which basically means that the spectrum of the input signal is multiplied by weighting coefficients calculated according to a weighting rule; their 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

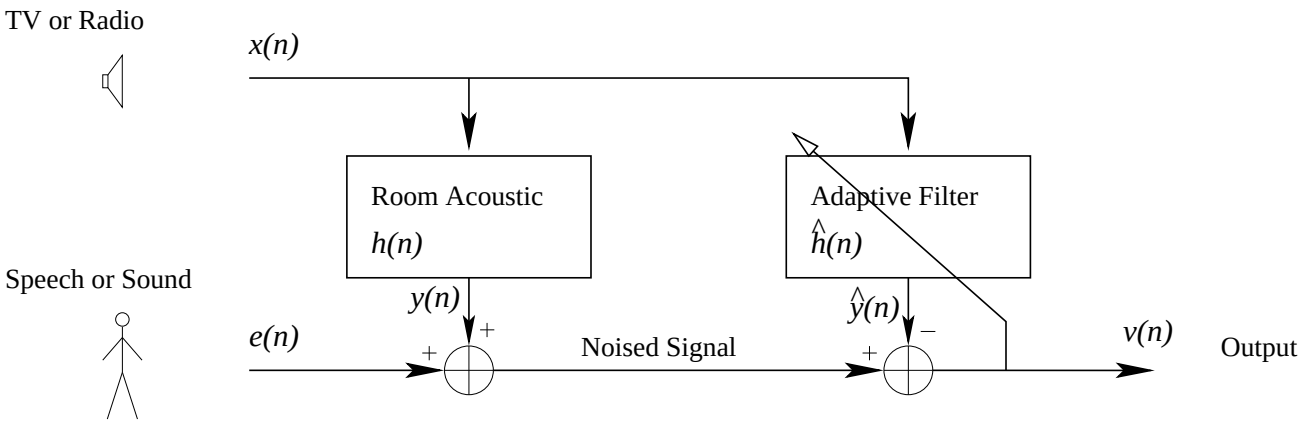
VII. NOISE SUPPRESSION EXPERIMENTS

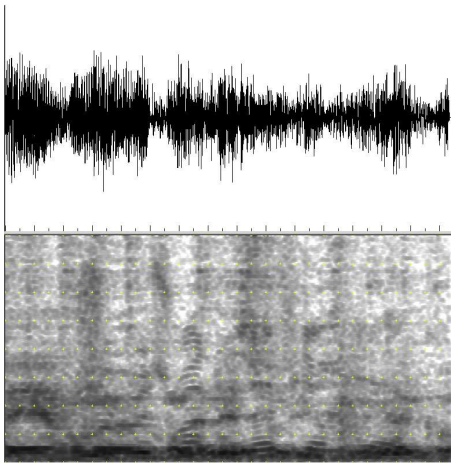
Two microphones were set in a room, the Reference Microphone in front of the Speaker System in order to record music or radio news (France-Info, a French radio broadcasting news all the day) and the Signal Microphone in order to record a French speaker uttering sentences in the room as shown on Figure 5. The two microphones are connected to the Audio Analysis System in charge of echo-cancellation; the resulting signal after echo-cancellation with or without post-filtering is then sent to the ASR and stored for further analysis. For this experiment the French speaker is standing in the center of the recording room, he is not facing the signal microphone. He has to speak with a normal voice level, the power level of the radio is set to be rather strong and then the SNR may be approximately 0 dB.

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 previous studies [10] may be used for this purpose. Echo-cancellation is operated in batch accorded to the reference and the addition of speech and noise. Hence, it is possible to proceed with the same signal using different settings of the echo-canceller. The results obtained with these two approaches are presented in the next section.

The spectrogram and the corresponding signal are displayed on Subfigure 6b in the case of the sentence "*J'ai besoin d'une infirmière*" and the corresponding noised signal at -6dB SNR on Subfigure 6a. After echo-cancellation the noise level remain quite significant as shown on Subfigure 6c but words are separated. The corresponding spectrogram shows that noise is present in all the frequency bands. By using post-filtering in conjunction with echo-cancellation, noise is low, as shown on Subfigure 6d, but the high frequencies of the original signal are attenuated. Harmonic components are not changed. It is then possible that speech recognition may be altered; it will be different according to the phonemes making up the sentence.

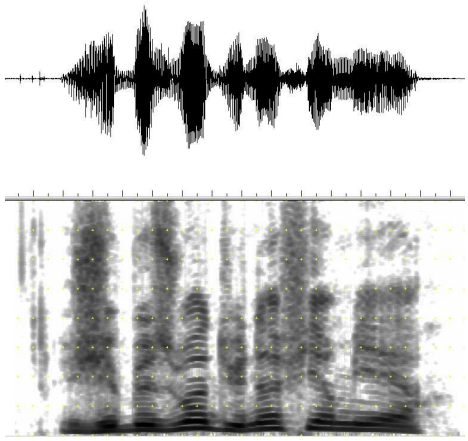
Figure 4. Block Diagram of Echo Cancellation System





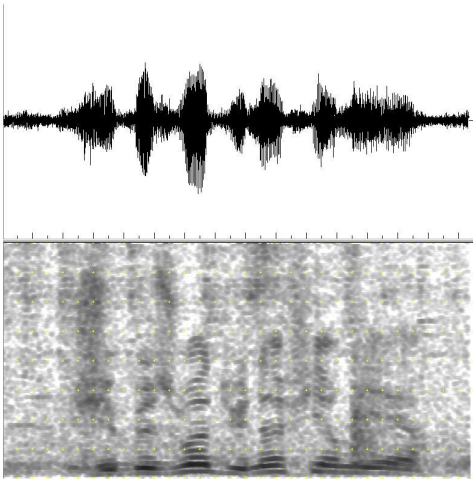
Subplot a: Noised sentence (SNR = -6 dB), *France-Info*

Best recognition hypothesis: "*Ferme les volets*"



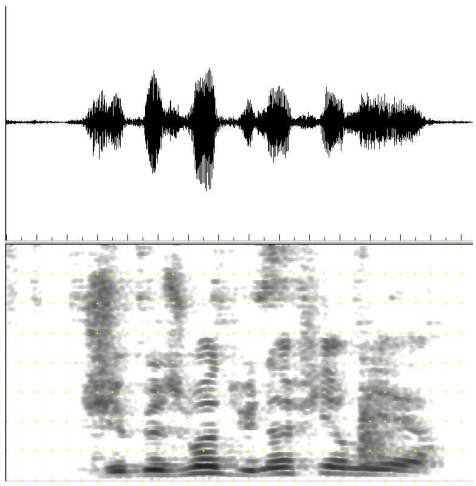
Subplot b: Original sentence

Uttered sentence: "*J'ai besoin d'une infirmière*"



Subplot c**:** Noised sentence after echo-cancellation

Best recognition hypothesis: "*C'est très mauvais infirmière*"



Subplot d: Echo-cancellation and post-filtering

Best recognition hypothesis: "*J'ai besoin d'une infirmière*"

Figure 6. Time representation and spectrogram of the signal: original, noised, after echo-cancellation and or without post-filtering (example sentence of the Normal/Distress corpus: "*J'ai besoin d'une infirmière*")

VIII. ECHO-CANCELLATION EVALUATION

The reference signal and the resulting noise in the room were recorded by 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 is then processed by the echo-cancellation system and the 126 sentences were extracted and sent to the ASR. This process is iterated a second time by the echo-cancellation system with post-filtering. The language model of the ASR was a medium vocabulary statistical system (9,958 words in French). This model is obtained by extraction of textual information from the Internet and from the French newspaper "Le Monde". Then, it is 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.

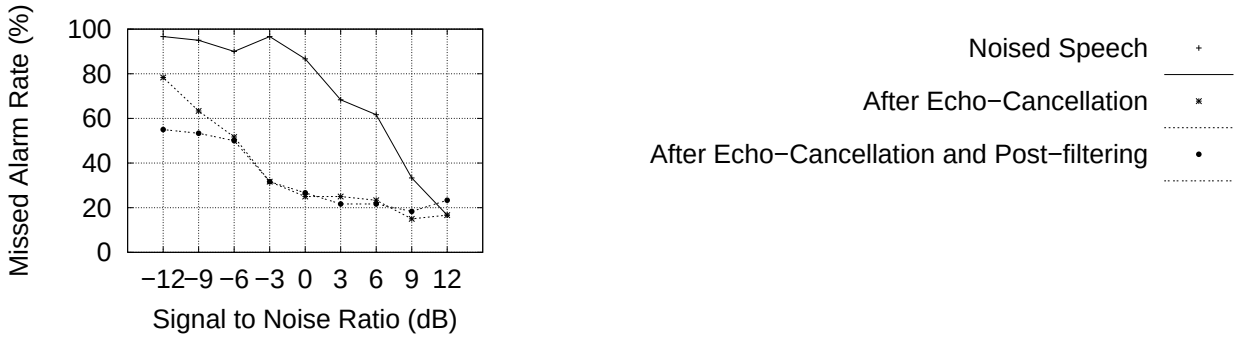


Figure 7. Missed Alarm Rate after Echo-Cancellation and Post-filtering as a function of the SNR

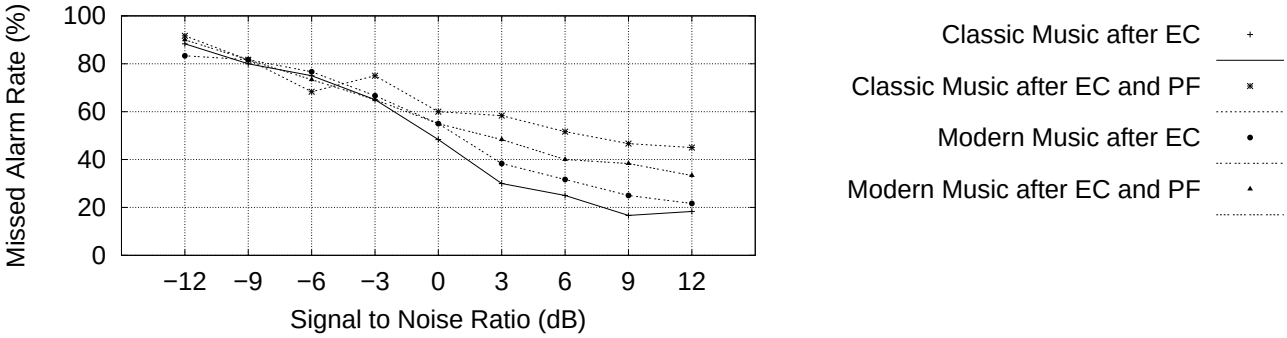


Figure 8. MAR with Music as Noise Source as a function of the SNR

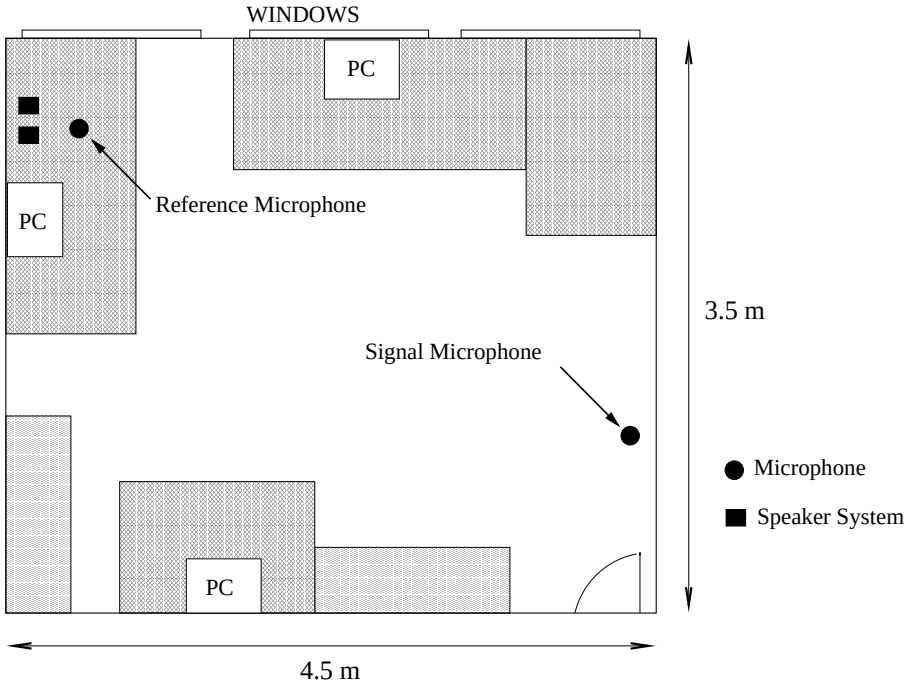


Figure 5. Setting of the microphones and speaker system in the recording room

In the absence of echo-cancellation, distress keywords are badly recognized, the MAR is fast increasing since +10 dB. The MAR curve is nearly flat between -3 dB and + 12 dB when echo-cancellation is processed, the post-filtering doesn't improve significantly speech recognition in this interval, the MAR is even greater at +12 dB. On the contrary, 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: classic music (The 3rd symphony opus 55 by Beethoven) and pop music (Artificial Animals Riding on Neverland by AaRON). Results are displayed on Figure 8, the error rate increases linearly with noise level. This kind of noises, and especially pop music, are more difficult to suppress because of the presence of large band sources like percussion instruments.

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 the speaker 2 or 3 times. The echo-cancellation was operated in real-time by the Audio System analysis. The level of the radio France-Info was set in order to achieve a 0 dB SNR level, each speaker was standing in the center 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 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.

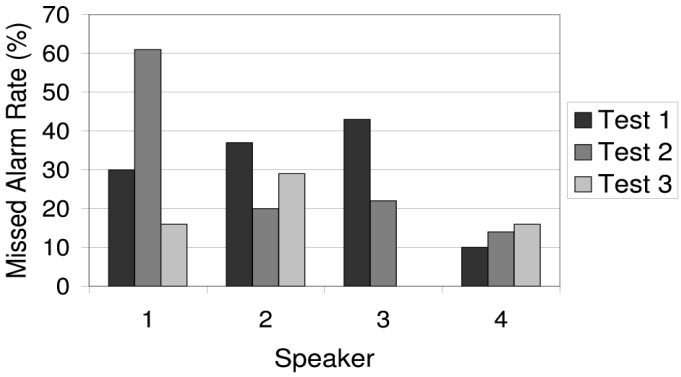


Figure 9. Missed Alarm Rate with Echo-Cancellation in Real-Time as a function of the speaker

IX. CONCLUSION

This paper presents the results of distress situation detection from speech in a smart room in the context of telemonitoring for elderly people at home. The first experiment presented was realized in a HIS flat where 10 speakers were involved and played a scenario. Our main requirement is the correct detection of a possible distress situation through keyword detection, without understanding the patient's conversation. Each speaker uttered about 45 sentences of the Normal/Distress speech corpus, along with sentences currently uttered during a telephone conversation. The Global Error Rate is 15.6%, the performance of the ASR may be improved in future studies by taking into account sounds emitted in spontaneous speech: tongue clicking, hesitatingly speaking and phonetical variants.

The second experiment is related to radio noise cancellation. Four speakers were involved in this experiment and uttered distress sentences during the listen of the news on France-Info. The Missed Alarm Rate was 27%. Some improvements must be added to the Echo-Cancellation method when the noise source is not speech but music.

Our future work will be the study of aging voice because most of the patients or elderly people at home are more than 70 years old. Moreover, the characteristics of their voice are very different and must be studied in order to detect distress or assisting requirement from the speech.

REFERENCES

[1] M. CHAN, D. ESTÈVE, C. ESCRIBA, E. CAMPO, "A review of smart homes-Present state and future challenges". Computer Method and Programs in Biomedicine, Vol.91(1), pp. 55-81, 2008.

[2] C. N. SCANAILL, S. CAREW, P. BARRALON, N. NOURY, D. LYONS, and G. M. LYONS, "A Review of Approaches to Mobility Telemonitoring of the Elderly in their Living Environment". Annals of Biomedical Engineering, Vol.34, pp. 547-563, 2006.

[3] G. LEBELLEGO, N. NOURY, G. VIRONE, M. MOUSSEAU, and J. DEMONGEOT, "A Model for the Measurement of Patient Activity in a Hospital Suite". Information Technology in Biomedicine, IEEE Transactions on, Vol.10(1), pp. 92-99, 2006.

[4] B. BOUCHARD, A. BOUZOUANE, and S. GIROUX, "A Smart Home Agent for Plan Recognition of Cognitively-Impaired Patients". Journal of Computers, Vol.1(5), pp. 35-62, 2006.

[5] M. STÄGER, P. LUKOWICZ, and G. TRÖSTER, "Power and accuracy tradeoffs in sound-based context recognition systems". Pervasive and Mobile Computing, Vol.3(3), pp. 300-327, 2007.

[6] J. C. WANG, H. P. LEE, J. F. WANG, and C. B. LIN, "Robust Environmental Sound Recognition for Home Automation". Automation Science and Engineering, IEEE Transactions on, Vol.5(1), pp. 25-31, 2008.

[7] M. VACHER, A. FLEURY, J.-F. SERIGNAT, N NOURY, and H. GLASSON, "Preliminary evaluation of speech/sound recognition for telemedicine application in a real environment". The 9thAnnual Conference of the International Speech Communication Association, INTERSPEECH'08 Proceedings, pp. 496-499, Brisbane, Australia, 2008.

[8] M. VACHER, J.-F. SERIGNAT, and S. CHAILLOL, "Sound Classification in a Smart Room Environment: an Approach using GMM and HMM Methods". Advances in Spoken Language Technology, SPED 2007 Proceedings, pp. 135-146, Iasi, Romania, 2007.

[9] J.-L. GAUVAIN, L.-F. LAMEL, M. ESKENAZI, "Design Considerations and Text Selection for BREF, a large French read-speech corpus", ICLSP, ICSLP'90 Proceedings, Kobe, Japan, pp. 1097–1100, 1990.

[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”, Lecture Notes in Computer Science, Artificial Intelligence, Text Speech and Dialogue, vol. 4188/2006, pp. 711-718, 2006.

[11] F. MICHAUT, M. BELLANGER, “Filtrage adaptatif : théorie et algorithmes”. Hermès Science Publications, Lavoisier, 2005.

[12] J.-S. SOO, K. K. PANG, “Multidelay block frequency domain adaptive filter”. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.38(2), pp. 373-376, 1990.

[13] J.-M. VALIN, “On adjusting the learning rate in frequency domain echo cancellation with double talk”. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.15(3), pp. 1030-1034, 2007.

[14] J.-M. VALIN, I. B. COLLINGS, “A new robust frequency domain echo canceller with closed-loop learning rate adaptation”. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’07 Proceedings, Vol.1, pp. 93-96, 2007.

[15] Y. EPHRAIM, D. MALAH, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator”. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.32(6), pp. 1109-1121, 1984.

[16] I. COHEN, B. BERDUGO, “Speech enhancement for non-stationary noise environments “. Signal Processing, Vol.81, pp. 2043-2418, 2001.

[17] S. GUSTAFSSON, R. MARTIN, P. JAX, P., VARY, “A psychoacoustic approach to combined acoustic echo cancellation and noise reduction”. IEEE Transactions on Speech and Audio Processing, Vol.10(5), pp. 245-256, 2004.

[18] S. V. VASEGHI, “Advanced Signal Processing and Digital Noise Reduction”. Chichester, U. K.: Wiley and Teubner, 1996.

[19] D. VAUFREYDAZ, C. BERGAMINI, J.-F. SERIGNAT, L. BESACIER, M. AKBAR, “A new methodology for speech corpora definition from Internet documents”. Language Resources and Evaluation Conference, LREC’2000 Proceedings, pp. 423-426, 2000.